

Validation of a Proposed Objective Assessment Tool for Ultrasound Image
Acquisition Utilizing the Focused Assessment With
Sonography for Trauma Examination

By
Markus T. Ziesmann

A thesis submitted to the Faculty of Graduate studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Surgery
University of Manitoba
Winnipeg

Copyright © 2013 Markus T. Ziesmann

Abstract

Introduction:

No protocol for the assessing ultrasound imaging skill has been validated. We sought to develop and validate an assessment protocol for ultrasound imaging for the Focused Assessment with Sonography for Trauma.

Methods:

Our assessment tool consisted of task checklists, a global rating scale, and hand-motion analysis and was developed by a modified Delphi technique. Novice and expert cohorts were recruited to perform a FAST exam on a volunteer for assessment under the protocol.

Results:

Experts scored higher on static image acquisition (11.58 of 16 versus 6.63, $p < 0.0001$), dynamic image acquisition (17.21 of 24 versus 11.08, $p = 0.0005$), and our global rating scale (29.79 of 40 versus 18.42, $p < 0.0001$); experts used fewer movements (263.0 movements versus 452.4, $p = 0.0216$) and a shorter path length than novices (60.097 m versus 32.777 m, $p = 0.0041$).

Conclusion:

Our protocol for assessing ultrasound imaging skill has criterion validity in assessing expertise and may lead to improvements to training and credentialing programs.

Acknowledgements

This project would not have been possible without the support of the Department of Surgery at the University of Manitoba, including my program directors Dr. Chris Andrew and Dr. Brendan McCarthy, as well as the Department Research Chairs Dr. Helmut Unruh and Dr. Richard Keijzer. Support and advice from Ms. Mary Brychka was always greatly appreciated. The Department of Surgery GFT Fund generously provided financial backing for this project.

My academic leave was made substantially more enjoyable by working with Dr. Sarah Steigerwald and Dr. Jennifer Metcalfe who were (and remain) excellent colleagues and friends.

I would also like to acknowledge the unwavering support and understanding provided by my family and friends, specifically including my parents Mrs. Joann and Dr. Manfred Ziesmann and my parents-in-law Mrs. Nancy and Mr. Jeff Gill. Above all, I would also like to thank my wife Jennifer for her support and commitment during long hours and stressful moments over the course of these studies.

Finally I would like to thank those directly contributing to this research project. Thanks to Mr. Blair Peters, our ultrasound model, to Dr. Chau Pham and Dr. David Kirschner for advice and recruitment, to Dr. Sarvesh Logsetty for data collection and design, to Dr. Atul Sharma for timely and valuable statistics consulting, and to Dr. Jason Park, Dr. Ashley Vergis, Dr. Andrew Kirkpatrick, and all the members of the Canadian Emergency Ultrasound Society for their contributions to the study design and Delphi process. Most of all, a tremendous thanks is owed to Dr. Lawrence Gillman for his mentorship and guidance on this project and others that we've worked on together; he has been an excellent example of scholarship, professionalism, and dedication to academia and is a terrific role model for aspiring academic surgeons in this province.

Table of Contents

Introduction.....	1
Defining Our Scope of Interest.....	1
Literature Review.....	5
History of Objective Technical Skills Assessment.....	5
Requisite Features of Useful Evaluative Tools.....	7
Global Rating Scales as Assessment Tools	11
Task Specific Checklists as Assessment Tools.....	14
Combined-Score Assessments	16
Hand Motion Analyses as Assessment Tools	19
Virtual Reality and High-Fidelity Simulation as Assessment Tools	22
Error Analysis as Assessment Tools.....	25
Ultrasound Skills Evaluation in the Literature.....	27
Conclusions and Preliminary Study Design Proposal	33
Methods.....	36
Development of Scoring Systems.....	36
Hand Motion Analysis Hardware and Software	39
Study Design and Enrolment	40
Data Scoring and Video Review.....	43
Statistical Analysis.....	44
Results.....	46
Introduction.....	46
FAST Image Acquisition Checklist.....	49
FAST Representative Image Checklist.....	52
Global Rating Scale of FAST Image Acquisition Checklist	54
Hand Motion Analysis	55
Inter-Rater Agreement	59
Logistic Regression Analysis.....	61
Discussion.....	71

Proposal, Design, and Methods	71
Data, Results, and Comparison to the Literature	74
Implications of Outcomes and Future Proposals	95
Conclusion	109
References	110
Appendices.....	124
Appendix A: FAST Image Acquisition Evaluation Checklist.....	124
Appendix B: FAST Representative Image Evaluation Checklist.....	125
Appendix C: Global Rating Scale of FAST Image Acquisition	126
Appendix D: Study Consent Form Signed by Participants.....	127
Appendix E: Pre-Participation Questionnaire.....	132

List of Tables

Table 1: Participant Demographics.....	46
Table 2: Hepatorenal Space Scores on FAST Image Acquisition Checklist.....	49
Table 3: Splenorenal Space Scores on FAST Image Acquisition Checklist.....	49
Table 4: Pelvis Space Scores on FAST Image Acquisition Checklist.....	50
Table 5: Pericardium Space Scores on FAST Image Acquisition Checklist.....	50
Table 6: Regional Score Summaries on FAST Image Acquisition Checklist.....	51
Table 7: Hepatorenal Space Scores on FAST Representative Image Checklist.....	52
Table 8: Splenorenal Space Scores on FAST Representative Image Checklist.....	52
Table 9: Pelvis Space Scores on FAST Representative Image Checklist.....	53
Table 10: Pericardium Space Scores on FAST Representative Image Checklist.....	53
Table 11: Regional Score Summaries on FAST Representative Image Checklist.....	53
Table 12: Outcome of Global Rating Scale of FAST Image Acquisition Scale as a Predictor of Expertise.....	54
Table 13: Path Length of Hand Motion by Novice and Expert FAST Performers.....	55
Table 14: Number of Movements to Complete FAST Exam.....	57
Table 15: Hand Motion Rate of Change Performance Observations.....	58
Table 16: Inter rater agreement on scoring of FAST Image Acquisition Checklist.....	59
Table 17: Inter rater agreement on scoring of FAST Representative Image Checklist.....	59
Table 18: Inter rater agreement on scoring Global Rating Scale of FAST Image Acquisition.....	60
Table 19: Univariate Logistic Regression Equations.....	61
Table 20: ROC Inflection Points for Global Rating Scale and Image Acquisition Checklist... ..	62
Table 21: Leave-One-Out Validation Model Prediction Errors.....	68
Table 22: Multivariate Logistic Regression Model of All Significant Univariate Scores.....	68
Table 23: Multivariate Logistic Regression Model of GRS and Image Acquisition Performance.....	70
Table 24: Kappa correlation and clinical interpretation from Viera et al.....	75
Table 25: Comparison of Task Checklist Findings Between Similar Ultrasound Skills Assessment Studies.....	76

Table 26: Comparison of Global Rating Scales Among Similar Ultrasound Skills Assessment Studies.....	78
Table 27: Intraclass Correlation of Gain Scores Divided by Class and Reviewer83
Table 28: Summary of Hand Motion Domain Validity Statistical Significance in the Reviewed Literature.....	86

List of Figures

Figure 1: Frequency of Ultrasound Usage Among Participants	47
Figure 2: Frequency of FAST Usage Among Participants	48
Figure 3: Total Path Length of Hand Movement by Novice and Expert Cohorts (mm)	56
Figure 4: Image Acquisition Checklist Model Receiver Operator Curve.....	63
Figure 5: Representative Image Checklist Model Receiver Operator Curve	64
Figure 6: Global Rating Scale Model Receiver Operating Curve	65
Figure 7: Total Path Length Model Receiver Operator Curve	66
Figure 8: Total Movements Model Receiver Operator Curve	67
Figure 9: Distribution of Total Gain Scores Within the Representative Image Checklist.....	82
Figure 10: Distribution of Total Gain Scores Within the Image Acquisition Checklist.....	83
Figure 11: Hand Motion Mapping of A Single Expert Participant’s FAST Performance.....	90
Figure 12: Hand Motion Mapping of A Single Novice Participant’s FAST Performance.....	91

Introduction and Defining Our Scope of Interest

Following proper technique is important in all medical procedures, but in few procedures is it *as* important as in ultrasonography. Ultrasonography utilizes the principle of measuring the reflections of sound waves as they are projected through tissues to generate visual representations of underlying structures. There are myriad benefits of ultrasound investigations and in many fields of medicine – for example imaging of the biliary tree¹, imaging of the pelvic organs², and assessing blood vessels³ – ultrasonography has become the first-choice of diagnostic imaging modality. Of particular note is that the recent interest in ultrasound has come largely from bedside or so-called “point of care” ultrasound investigations performed by non-radiologist physicians and even non-physicians⁴; a recent American survey found that 71% of urban emergency departments are capable of performing point of care ultrasound investigations⁵.

What makes point of care ultrasound a unique test in medicine is that it is one of the rare diagnostic tests performed where the physician is responsible for both capturing *and* interpreting the data. Nearly all other diagnostic modalities employed by physicians place the practitioner in the role of interpreter *or* data acquirer; a physician may interpret an x-ray but is not routinely responsible for capturing the actual image in an example of the former, or may obtain a sample of cerebrospinal fluid but does not personally culture the sample in the latter. Ultrasonography requires the physician to employ the device to generate images and interpret those images, usually in real-time with live, dynamic pictures.

The growth of point of care ultrasonography has greatly expanded the role of ultrasound in medicine in general, but particularly in the management of acutely unstable or critically ill patients^{6,7}. While ultrasound is now used to assess urgent or emergent medical problems

including elevated intracranial pressure⁸, pneumothorax⁹, pulmonary emboli¹⁰, and intravascular volume status¹¹, its most common and well-studied emergency bedside use remains the Focused Assessment with Sonography for Trauma (FAST)¹² examination. This exam is a rapid point of care test performed by the treating physician managing an unstable and critically injured trauma patient. It asks a single question: is there evidence of fluid in regions where there should not be? The FAST exam encompasses an ultrasound investigation of the right upper quadrant of the abdomen, left upper quadrant, pelvis, and pericardium. The test can be successfully performed in 2-4 minutes¹³⁻¹⁵ by an experienced practitioner and a positive test may take as little as 19 second to recognize¹⁶ with imaging of a single region¹³.

Should an individual lack the capacity to either generate or interpret the ultrasound images when attempting to perform the FAST exam, the test itself becomes useless. When ultrasound images are of poor quality, diagnostic capability is decreased¹⁷. This is of particular concern given the target population of the FAST exam: hemodynamically unstable major trauma victims whose clinical course demands rapid and accurate decision-making. While there are a multitude of studies which are able to describe the clinical utility of the FAST exam and address the interpretive outcomes of the test¹⁸⁻²⁰, no study to date has attempted to answer the question of how we can objectively assess the technical aspects of image acquisition in FAST sonography. We do, however, see evidence that a physician's personal skill level with FAST impacts their outcomes, as more experienced physicians are able to achieve higher positive predictive values on the exam. This affirms our belief that, as a user-dependent skill, some objective assessment is in order²¹.

The lack of published objective imaging technique assessment guidelines for FAST ultrasonography in the literature prompted us to develop our own objective imaging technique assessment tool.

In reality, this field of research is not unique to FAST or ultrasound. Every field of medicine includes some manner of procedures to be performed in a clinical setting, and as such, every subgroup of medical trainees requires training in, and assessment of, their ability to perform procedures. While some fields such as surgery or anesthesia have a procedure-focused training curriculum, other fields perform medical procedures on patients infrequently. Procedural training, regardless of its asymmetric distribution amongst fields of practise, presents a unique problem for the medical educator: on what basis does an evaluator determine that a candidate has successfully mastered the requisite procedures to be learned in their field of study?

The question is not new, but has not been adequately addressed in the existing literature. The majority of published studies on the subject of procedural evaluation have as their endpoint a clinical outcome. For example, a study on biopsy technique by trainees may measure the success rate of achieving a viable sample²², or a study of plaster cast application technique may measure the rate of achieving anatomic reductions in bone fragments²³. However, the use of clinical endpoints in the assessment of technical capabilities may fail to detect significant technical deficiencies within a procedure of interest.

While the clinical end points may be carefully defined to *imply* better technique, no clinical endpoint can truly determine if the technical performance was adequate. In the previous example of pathology specimen acquisition as an endpoint for biopsy technique, we may demonstrate adequacy of sample harvesting while failing to demonstrate whether adequate analgesia was provided, if adequate hemostasis was achieved, or if sterile technique was

employed. Successfully meeting the clinical endpoint may imply that the user has obtained a degree of technical proficiency, but it is incapable of assessing the safety and consistency of the methodology employed by the trainee. Furthermore, when a clinical endpoint is *not* reached, the educator, researcher, or student is not provided information as to why. Alternatively, an observational study powered to recognize *all* significant outcomes including measures of morbidity would require the exposure of patients to substantial risk, as any trainee failing to reach certification standards would by definition have had poor outcomes in one or more defined measures of morbidity or poor results in the primary clinical outcome. Despite the need for technical proficiency assessments, 90% of American surgical fellowship programs have no formal skills assessment as part of their training requirements²⁴.

Literature Review: History of Objective Technical Skills Assessment

The modern era of medical training with reduced training hours and case volumes by many residents^{25,26} has spurred a recent interest in objective technical assessment²⁷⁻²⁹. It has been recognized by the Royal College of Physicians and Surgeons of Canada that a movement towards a competency-based education system and away from a time-based education system is necessary³⁰. This mirrors the Accreditation Council for Graduate Medical Education's "Outcome Project"³¹ in the United States of America. Defining measures of competence requires validated tools to assess performance in an objective manner, and tools must be developed and validated that are capable of assessing each major skill set required of trainees.

Historically, evaluations of technical skill took the form of subjective evaluations completed by supervising staff³². Because technical skill assessments in this manner usually rely on retrospective opinions, are unstandardized, unvalidated, and subject to bias, their utility in a competency-based assessment program are limited^{32,33}. The common In Training Evaluation Report (ITER) used by many faculties has been shown to be subject to various biases depending on the trainees being evaluated, including a bias towards excessive leniency³⁴ and a lack of correlation with performance on Objective Structured Clinical Examinations (OSCEs)³⁵. Additionally, as many programs generate an evaluation comprised of the summative feedback from multiple faculty members, there is a potential for bias to be introduced when feedback does not originate equally from all potential sources³⁶.

The solution to this problem has been to couple educational programs with multiple other requirements for graduation in addition to preceptor feedback assessments. In his seminal paper on the subject of technical skills assessment, Richard Reznick noted that the accepted standards

for technical assessment include some combination of self-reported experience, evaluations by a preceptor with or without criteria, simulation or modeled assessments, and video-taped real world performance³². Trainees have traditionally been required to complete a specified number of “blocks” of training over a specified number of years in addition to passing a knowledge exam. Unfortunately, little evidence exists to link the time spent training or the number of procedures performed to competence with complex tasks. Analysis of a large prospective cohort of surgical endoscopists for example found no correlation between a surgeon’s experience level and their success with the procedure, or with their complication rates³⁷. Looking at a complex skill similar to our target procedure, FAST ultrasound, another evaluation compared cardiology fellows’ echocardiography skills in both interpretive and scanning domains and sought to correlate OSCE performance with their experience (measured both in duration of training and personal experience performing echocardiograms). Despite all fellows having received subjective evaluations indicating they met or exceeded training expectations, more than half of the cohort failed the interpretive test, the entire cohort produced images graded as “barely adequate” and poor correlation was found between OSCE scores and both measures of experience³⁸. The combination of subjective evaluations with minimum experience quotas is insufficient to allow us to declare an applicant “competent” and therefore some other objective measure is needed.

Literature Review: Requisite Features of Useful Evaluative Tools

To be a useful tool, any assessment modality needs to be valid, reliable, and feasible^{32,39}. The latter component is relatively easy to assess. Feasibility requires a process of evaluation to be safe, simple, well tolerated, and inexpensive, without requiring arduous training on the part of the evaluator³⁹. Our proposed FAST imaging assessment must meet each of these goals.

Validity and reliability come in multiple formats which must be independently addressed. In order to develop a high quality protocol we must consider different variations of validity and reliability to ensure our ultimate construct meets our needs.

Reliability overlaps somewhat with measures of validity, but in general can be summarized as a high level of agreement in scores from multiple tests of the same subject. Test-retest reliability measures the ability of the test function to deliver consistent results when the variable studied has not changed; without an intervention, an ideal test would deliver identical scores when repeated on the same subject⁴⁰. Inter-rater reliability compares the consistency of scoring from different judges on the same performance; an ideal test again would produce identical results of the same performance as evaluated by different reviewers⁴⁰. Internal consistency is another form of reliability which measures the agreement between multiple independent assessors of a single variable; for example, in a multiple choice exam, the scores on all questions on the same subject would be expected to strongly correlate in an internally consistent test⁴⁰. The development of our scoring protocol should therefore include multiple assessments of the same skill set, and be tested repeatedly and by multiple reviewers to ensure that we demonstrate test-retest, inter-rater, and internal consistency forms of reliability.

Validity, like reliability, has multiple specific definitions which represent different aspects of our test. *Construct validity* is the premise that the measurement of one variable may serve as a logical proxy for another facet of assessment; it is similar to internal consistency (a reliability measure) as described above⁴¹⁻⁴³. Construct validity is derived from two constituent components. *Convergent validity* is present when related variables in a logical construct are found to correlate positively^{43,44}. For example, in a hypothetical study that demonstrates worsening exercise tolerance in a population with heart failure, convergent validity would be demonstrated by the increased reporting frequency of dyspnea. *Discriminant validity* is found when unrelated variables in a logical construct are found *not* to correlate^{43,45}. When both convergent and discriminant findings are present, a test can be said to have construct validity. A test with construct validity may allow us to assess the underlying skill set or traits of testees that may otherwise be difficult to measure. Furthermore, expertise may in fact become a proxy variable in the determination of construct validity; a recognized expert should be expected to perform better at a technical skill than a known novice. This trait can be used to validate assessment modalities.

Content validity identifies whether a model encompasses all important elements of the subject being assessed^{42,43}. A model of ultrasound certification based on interpretation of images without image acquisition skills assessment for example, would lack content validity because the assessment fails to capture a significant component of the skill. It is important to distinguish content validity from *face validity*. Face validity is the superficial appearance of an assessment modality capturing all facets of the test subject, or the appearance that an assessment modality “makes sense” to the observer; whereas face validity is based on subjective opinion, content validity requires statistical evidence to be present^{43,46}. A test may have face validity without

content validity. For example, one may presume that academic grades achieved by applicants to surgical residencies would predict clinical performance of the trainees; in fact this has not been shown to be the case, despite the face validity of the presumption⁴⁷.

Criterion validity is an assessment comparing our target assessment to proven measures of the target variables, usually using some accepted “gold standard” validated assessment. It is further divided into *predictive validity* and *concurrent validity*. Predictive validity describes the test’s ability to predict future outcomes with the present measurements⁴³. Predictive validity, for example, would be represented by the ability to predict real-world performance from test results in trainees. Concurrent validity describes the agreement between the test scores and actual performance of the studied technique⁴³. When both concurrent and predictive validity are present – that is, our proposed test correlates with both present and future assessments against a gold standard – we accept that there is criterion validity.

Achieving our goals in proving various types of validity will require careful planning. The use of multiple modalities of assessment for a single criteria, in addition to demonstrating internal consistency and reliability, can allow us to assess construct validity including both convergent and divergent validity. Our proposed model requires an as realistic-as-reasonably-possible demonstration of skills, ideally based on the assessment of the performance of an actual clinical investigation or intervention to ensure content validity. Finally, we must compare our assessment against a proven standard of assessment to demonstrate criterion validity. Our proposal must therefore, in summary, include multiple formats of assessment to be scored by multiple assessors, be based around a realistic performance of the assessed skill rather than a low fidelity simulation, and must be compared against a known standard of assessment in order to demonstrate each important aspect of validity.

Accepting that a hypothetical ideal technical skills assessment would be valid across all definitions, we are left with the task of designing such a test for technical proficiency in image acquisition during FAST ultrasonography. In their 2001 review paper on surgical skills assessment, Darzi and Mackay argued that in addition to being valid, a technical capabilities assessment protocol must assess three criteria³³. A technical assessment should consider:

1. Surgical judgement
2. Knowledge (of the procedural steps)
3. Dexterity

We can apply those principles to ultrasonography with slight modification. The analog might be represented as:

1. Image optimization
2. Knowledge (of anatomic landmarks)
3. Probe handling

These principles are in keeping with our outline for testing described above. Image optimization, probe handling and knowledge are all inter-related components which independently will contribute to a good overall ultrasound imaging performance. In this sense, an assessment of each of these three arms will contribute to internal consistency and construct validity in our assessment. An assessment derived from these considerations still requires validation after rigorous testing. Multiple different methods of technical skill assessment have been proposed including rating scales, checklists, hand motion analysis, virtual reality, and error evaluation. To establish criterion validity, our assessment should be based on established models from the literature.

Literature Review: Global Rating Scales As Assessment Tools

A global rating scale (GRS) is a task non-specific objective scoring system using a Likert scale as a manner of assessment^{39,48}. The underlying principle behind a global rating scale is that a set of technical proficiencies deemed necessary to a field of practise are assessed, but the scale must not be specific to a single task. It should therefore measure general proficiencies within a skill set and not competence to perform a single procedure. One advantage of a global versus a specific scale is that a GRS may be applied to multiple related tasks without the onerous need to develop and validate multiple task-specific scales⁴⁸. Global rating scales may vary in the number of and description of domains assessed, with specific anchoring of the Likert scale to performance objectives.

One benefit of the use of global rating scales is their wide ranging use with proven validity for many tasks. GRS models of technical skills have been validated in the evaluation of orthopedic arthroscopic procedures, both task specific^{49,50} and general⁵¹, in addition to open surgical procedures⁵². In general surgery, GRS scales have been developed and validated for specific procedures such as laparoscopic cholecystectomy and appendectomy⁵³⁻⁵⁶ and laparoscopic fundoplication and anterior resections⁵⁷. Some studies have demonstrated a high degree of inter-task correlation using either a single GRS to assess multiple procedures³⁹ or a GRS combined with other measures⁵⁸. Global ratings scales, usually combined with other methods of evaluation, have also been validated in the evaluations of surgical skills in cardiac surgery⁵⁹, ophthalmology⁶⁰, and gynecology^{61,62}.

Our own question – how to assess technical proficiency with ultrasonography – requires the development of a GRS for non-surgical domains. Surgical domains in rating scales include

elements such as “respect for tissue,” “knowledge of instruments,” and “use of assistants” which are not reflective of techniques used in most ultrasound procedures. It is notable therefore that GRS assessments have been developed and validated in non-surgical realms including epidural catheter insertion⁶³, endotracheal intubation⁶⁴ and gastrointestinal endoscopy skills⁶⁵. It is clear that the use of a global rating scale is a well-proven concept with wide-ranging applicability.

From reviewing these studies, however, a GRS scale alone may not be the best test for our assessment. By design, a GRS is a universally applicable tool measuring skills within a broad domain; it does not measure the completion of specific tasks during a procedure. This is apparent when we see high degrees of inter-task agreement in scales used to assess a heterogeneous collection of tasks^{39,58}. For example, a single GRS proposal used by Doyle et al. to assess trainees performing fifteen different surgical procedures including such dissimilar items as “[arteriovenous] fistula creation,” “[thoracoscopic] evacuation of empyema,” “hemithyroidectomy,” and “proctocolectomy,” found that across procedures, scores were predicted strongly by post graduate training year³⁹. Intuitively, we are left to wonder if it makes sense that individual trainees should perform similarly in all procedures despite the overwhelming differences between them. In other words, is it reasonable to expect a high score for hemithyroidectomy performance simply because a resident achieved a high score in proctocolectomy performance?

One explanation may be that the exceptional trainees are exceptional at everything. Another explanation however is that measurement of specific-task accomplishment would provide some higher degree of discriminatory power that is lacking in a global rating scale. For example, a trainee scoring highly on the “knowledge of instruments” domain of a general surgery rating scale would be expected to score highly for both a laparoscopic cholecystectomy

(a common “novice level” surgical procedure for trainees) and a laparoscopic anterior resection (a much more “advanced level” surgical procedure) despite tremendous differences in the performance of these procedures. Residents who have mastered domains of simple skills are therefore at an inherent *scoring* advantage when performance of a more complex task is assessed, regardless of whether knowledge or technical competence in performing the advanced task is satisfactory, giving a falsely elevated score. Overlapping skills do not imply competence in a specialized task, as can be observed in the impressive technical ability of teenage video-gamers using laparoscopy simulators⁶⁶. Some task-specific domains may be necessary in our own scoring system, particularly due to the widespread use of ultrasound in many clinical domains. To raise the question in our own field of interest: is a trainee familiar with ultrasound gel application and gain adjustment from their experience with central venous catheter insertion by default an adept FAST sonographer? Of course the answer is “no.” It is clear that while a global rating scale may play a role in our assessment technique, some task-specific criteria are also necessary.

Literature Review: Task Specific Checklists as Assessment Tools

A task-specific checklist measures the achievement of specific steps in the completion of one singular procedure. While global rating scales may be designed to assess proficiency in multiple related skills, a new task-specific checklist must be developed and validated for assessing each different procedure. For example, laparoscopic appendectomy and cholecystectomy may be assessed by the *same* GRS, but require two *different* task checklists. Task-specific checklists must therefore be individually tested and validated for each task under consideration. The most common configuration is to create a detailed list of important steps in the studied procedure and to score the test as a binary “completed” or “not completed” sum. A completed item is scored as a “1” and an incomplete item is scored as a “0.” This is the classic method of scoring medical OSCEs (Objective Structured Clinical Examinations)⁶⁷⁻⁶⁹. Some task-specific lists however blur the distinction in design between a true checklist and a GRS-style scale, incorporating a “yes/no” aspect and a quality aspect (for example, scoring 0 for “not performed,” 1 for “performed late/poorly,” or 2 for “performed well and rapidly”)^{70,71} or even utilizing a task-specific five point Likert scale^{72,73}. The key distinction remains however, that a task-specific score should unequivocally apply to that single procedure, regardless of how it is measured.

Assessments utilizing task-specific checklists alone are uncommon because the completion of a task does not provide the reviewer information on the quality of task completion. The circumstances where task checklists are the exclusive evaluative tool are often non-procedural technical assessments; for example, the evaluation of resuscitation efforts^{70,71}. In these scenarios completion of a task is more critical than the means in which it was completed; for example, when trainees score points for “secures the airway” in a resuscitation scenario, it is

irrelevant if the view on laryngoscopy was good or bad, or if the airway was secured with direct laryngoscopy or the use of a gum elastic bougie, so long as the task was successfully and rapidly completed.

Literature Review: Combined-Score Assessments

When compared against one another as measures of the same task, global rating scales and task checklists have both been shown to be valid measures of skill. When asked which format of assessment is the most appropriate and most closely mirrors the subjective opinions of an evaluator, raters choose checklists over rating scales⁴⁸. When subjective preference is ignored however, global rating scales demonstrate improved inter-test reliability and improved construct and concurrent validity compared to checklists^{68,74}. Checklists in comparison demonstrate higher inter-rater reliability⁶⁸. It has also been observed that checklist scores in a true expert may paradoxically decrease on a clinical assessment, as expert-level diagnostic skills may not rely on the systematic approach necessary for a high score on many OSCE assessments, despite accuracy of the ultimate diagnosis⁶⁹. As both tests demonstrate unique properties, there is little consensus that one test is superior to the other.

For most technical assessments we require an assessment of quality as well as assurance that all critical tasks have been completed. A common practise therefore has been to combine a task-specific score (a step-by-step assessment) with a global rating scale (a quality of technique assessment). The model combining a task checklist with a GRS and an overall pass or fail grade was branded the Objective Structured Assessment of Technical Skill (OSATS) by Martin et al. in 1997⁵⁸. They hypothesized that by combining the two modalities of evaluation they would find moderate to strong correlations between them and also find equivalent scoring on bench versus live animal assessments of residents' surgical skills. Using a multivariate analysis of variance (MANOVA) assessment, the researchers found that 30% of a trainee's performance during objective assessment was attributable to level of residency training. Thus, the OSATS model has good construct validity as a measure of technical skill.

Since Martin et al. presented their combined task-specific and GRS methodology, new scoring systems based on theirs have been developed, tested, and validated in many fields of training. While Martin proposed the use of the OSATS as a means of assessing performance on bench models, assessments using simulation, animal models, and live surgery have also been validated. OSATS-style tests utilizing GRS and task checklists adapted to different realms and specific tasks have been validated for orthopedic surgery^{50,52}, general surgery^{54-58,72,75-77}, cardiac surgery⁵⁹, otolaryngology⁷⁸, and gynecologic surgery^{61,62}. Like task checklists and global rating scales, OSATS-style systems have also been utilized in non-surgical realms to assess anesthesia procedures^{63,64}, resuscitation efforts⁷⁹, and operative dictations^{80,81}. In many ways the OSATS has become the accepted “gold standard” of technical assessment due to its history as a well studied, widely validated, and diversely applicable assessment method. When studying other modalities of assessment, criterion validity is most frequently proven by comparison with synchronous OSATS test^{49,82,83}. Interestingly, the OSATS itself when compared to the gold standard of the era – the subjective assessment of trainees by preceptors – failed to demonstrate criterion validity, which is a finding consistent with the opinion that the historic standard was indeed flawed⁷⁵.

The primary criteria for “expertise” or “proficiency” in most checklist, rating scale, and OSATS validation studies has been year of training; an attending physician is defined to be more proficient than a senior resident, who is more proficient than a junior resident (and thus they should score, respectively, highly, moderately, and poorly on the same OSATS assessment)^{58,84,85}. Under such a definition it could be proposed that the OSATS is a measure of some proxy for technical proficiency rather than proficiency itself; under those definitions for example, a test simply measuring “years of experience” would provide the expected results,

though would obviously fail to meet face validity as an objective measure. While increased patient volumes in a surgical practise have been shown to improve clinical outcomes⁸⁶⁻⁹⁰, the number of repetitions performed during training does not show a robust relationship⁹¹. It is important to note, therefore, that not every OSATS validation study has demonstrated universally satisfying results. While there is a publication bias that minimizes the incidence of negative studies being published, the existing literature does demonstrate OSATS validation attempts which did not show a difference between cohorts. Swift et al. attempted to validate the OSATS using three cohorts – attending gynecologists, senior residents, and junior residents – participating in ten skills stations using bench models and simulations. In only six of the ten stations did they demonstrate significant and expected results consistent with experience; two stations showed insignificant differences between groups and one station showed a reversed trend with the junior residents scoring highest⁶¹. This underscores the key points that the results yielded by OSATS are dependent on good study design and each scoring system must be validated before any clinical application can be of merit.

Literature Review: Hand Motion Analyses as Assessment Tools

Other modalities of skills assessment have been proposed outside of pen-and-paper assessments such as rating scales, task checklists and OSATS combinations. Hand motion analysis is one modality which is particularly interesting in that it is a purely objective assessment tool. While a rating scale is designed to maximize objectivity, the inter-rater agreement will never reach perfection. If we could measure and record the movements of a participant's hands, no subjectivity need enter the equation and inter-rater agreement (that is, the hypothetical measurement of one performance with multiple hand-motion devices simultaneously) should approach perfection.

Hand motion devices require a participant to wear small electromagnets on their hands while a magnetic field generator and computer software measure the movements of the magnets (and thus, the trainee's hands) in three dimensional space⁹². Recording from the hand motion sensors can include time, total path length of travel, number of movements, velocity, and jerk. In theory, hand motion analysis should provide an impartial and unbiased assessment that assesses the efficiency of a testee at completing a task; an expert should be expected to complete a task in less time, with fewer movements, and a shorter path length than a less skilled trainee.

Evaluation using hand motion devices has been validated in multiple fields; unlike the OSATS or other scoring systems, hand-motion analysis is by nature only useful in procedural fields and is naturally unhelpful in scoring tasks such as resuscitation of a critically ill patient. Validated uses of hand motion analysis have been demonstrated in arthroscopic orthopedic surgery^{49,93-96}, laparoscopic general surgery^{56,76,97}, open general surgery^{98,99}, regional anesthesia^{92,100}, and ophthalmologic microsurgery⁶⁰. In most cases and with few outliers, all

three components – time of procedure, number of movements to complete a task, and the total combined path length of both hands – have reached clinical significance as discriminatory tests between cohorts of trainees. When compared against an accepted method of evaluation such as an OSATS-style protocol, hand motion analysis provides an instant, objective, repeatable and valid means of assessing technical skills that makes it a valuable tool for any skills assessor^{49,56,60,76,93–98}.

The detriments of hand motion analysis relate to both the test itself and the information it provides. While hand motion studies do provide objective and quantitative measures of performance, they provide less qualitative information than a task-specific checklist or global rating scale. While a task-checklist at least accounts for the steps being accomplished and can therefore imply a level of quality (albeit without an explicit *grading* of quality) hand motion analysis fails to provide any quality assessment or task-completion evidence. A trainee undergoing hand motion assessment may fail to complete the procedure but the measures of time, movements, and path length will not reflect this outcome in any way, and an assessor reviewing the data will also not be able to provide the trainee with useful feedback on specific points of failure or areas to focus future practise. Furthermore, while technically possible, hand motion devices are difficult to utilize in a live-operating environment and have largely been confined to use on bench models, while GRS and checklist assessments are easily applied to live operating performance⁴⁹. Hand motion can therefore not be used as an exclusive means of assessing skill, but may be a useful tool when combined with another modality of assessment^{96,98,101}.

Further concerns about hand motion as a modality for objective assessment pertain to the devices themselves. Motion capture devices require specialized hardware and software which often provide raw data requiring further statistical analysis before it becomes interpretable to the average clinician or educator. This mandates the purchase of expensive specialized equipment, both hardware and software, and statistical analysis training (or a budget for consultation) which may make hand motion assessments unavailable in many environments. Because many trainees experience at least some portion of their training in non-tertiary care facilities, we have to ask if it is reasonable to require all training sites to procure, utilize, and maintain complex hand motion analysis devices in order to evaluate trainees.

Literature Review: Virtual Reality and High-Fidelity Simulation as Assessment Tools

A similar but distinct objective assessment proposal includes the use of virtual reality devices to assess performance. Such devices rely on the proprietary assessment metrics built in to high fidelity simulators. Simulation technology has advanced significantly over recent decades to the point where mannequins exist that allow surgical procedures to be practised¹⁰². However, these devices often lack the fidelity or realism to act as a true “virtual reality” device. Most virtual reality (VR) systems employ computer-screen based procedures and are therefore ideal for tasks such as laparoscopic or endoscopic training.

Various models of virtual reality simulators have come onto the market since one of the most popular early models, the MIST VR¹⁰³. Validation of VR metrics includes an assessment of multiple parameters. While individual products on the market vary, in general the devices combine the hand-motion assessment of stand-alone electromagnetic devices with a simulated procedural task, allowing assessment in a safe environment of domains such as tissue destruction, forces used, blood loss, and objective success/failure in completing the tasks. Many studies utilizing simulation technology have a goal of proving that simulation training improves clinical performance in live operating theaters^{104–108}, an end-point which is not necessarily in keeping with our own goal of technical skills assessment. Still, the use of simulation scoring metrics as assessment tools has been successfully attempted by some authors. Gallagher et al. used laparoscopic simulation to compare a cohort of novices, an intermediate cohort, and an expert cohort and found that in all parameters – time, error, left hand movements, right hand movements, amount of cautery used, and consistency – the expert cohort outperformed the others^{109–111}. Scott et al. used a different simulator to assess skills, combining the simulator scores with an OSATS-style assessment, and found that the simulator metrics suggested

improved performance in all five simulated tests based on time to completion⁸³. Other studies have concluded that simulation can differentiate novices from experts in domains of time to complete tests, instrument-induced trauma, errors, clip placement, blood loss, tissue stretch, economy of motion, incision accuracy, camera path length, and instrument misses^{7,112-117}.

In addition to laparoscopic and endoscopic procedures, the recent development of open surgical simulators has also demonstrated validity in skills assessment. Such devices have shown that experts perform open procedures faster than trainees, with fewer deviations from predetermined steps in prostate surgery using a prototype device known as SimPraxis¹¹⁸. Prototypes have also been proposed or developed for a neurosurgical simulator (NeuroSim)¹¹⁹, a simulator to assess operative exposure¹²⁰, fracture reduction and amputation¹²¹, open inguinal hernia repair¹²², and open cardiac surgery¹²³. It is clear that while data for open surgical simulation is lacking today, this paucity will be rectified over the next decade.

The variety of tests validated by existing simulators demonstrates a major flaw of virtual reality metrics: there is inconsistency in the measures, algorithms, and designs of the devices on the market. While one machine may give metrics for path length and time, which one would expect to be easily reproducible and consistent between devices, another may use a proprietary algorithm to calculate “economy of motion,” estimated “tissue stretch” or “errors.” Even when two devices purport to measure the same criteria different devices may yield different results and machine specifications are inconsistent¹²⁴. Because there is no standard in the industry, any validation of a single device or study design contributes little to the overall body of literature in medical education as most other facilities will lack the ability to conduct the same test on the same device. Further, like hand motion analysis devices, high fidelity simulators are expensive

and it is impractical to suggest that all objective assessments must be completed on such a device.

Novel simulation technology has allowed the development of high-fidelity non-endoscopic ultrasound simulators¹²⁵⁻¹²⁷. These devices replicate an ultrasound device closely with probes, buttons, and image quality copied as closely as possible from fully functional ultrasound devices. The benefits of such devices is the so-called “pathology on demand” aspect which, as with other surgical simulators, allows a trainee or trainer to assess performance of the assessment of specific medical conditions rather than generic skills alone. The use of these devices has been shown to be valid in echocardiography¹²⁸ and obstetrical ultrasounds, during which trainees took longer to measure fetal size and did so with a higher variance than experts¹²⁹. In one trial, comparison of performance of residents trained in ultrasound on a simulator versus those who trained on live human volunteers was unable to detect a measurable performance difference in the two cohorts¹³⁰. While no published studies (as of yet) have utilized ultrasound simulators to measure the same metrics as other surgical simulators – namely hand motion, errors, path length, or other complex metrics – the ultrasound simulation field represents a novel and high-potential field for future study.

Literature Review: Error Analysis as an Assessment Tool

One final proposal for assessing performance is to ask trainees to evaluate a task at various stages of its completion and determine if the steps have been performed correctly or not. This method of assessment has been largely pioneered and studied by a single group of researchers, notably Simon Bann and colleagues. The underlying principle is that medical errors fall into multiple categories including slips (distraction failures), lapses (memory failures), and knowledge mistakes (a lack of knowledge of the procedure). Senior residents are thus more prone to slips and lapses because they generally have adequate knowledge, whereas junior residents are more prone to knowledge mistakes⁸². Bann and colleagues hypothesized that the ability to detect errors would correlate strongly with surgical performance; by creating a number of bench models of surgical procedures, some with errors by design, and asking trainees to evaluate the models for evidence of errors, they were able to create a scoring system for error assessment. This was subsequently correlated against an OSATS-style scoring system assessing performance of a technical skill. The findings suggested that error detection is a valid and reliable measure of technical proficiency, as the error scores correlated strongly and significantly with OSATS scores^{82,131,132}.

The concept of error analysis is indeed novel, but it has failed to garner much attention in the overall medical education community. One problem with error-based assessments is that they depend very much on a binary error; a suture may not be square, or a clip may be applied across another clip for example. Error analysis of most other tasks, including ultrasound, would be difficult as errors are far more subjective. For example: is a clear static image of the pericardium with well-adjusted gain and the cardiac structures clearly visible an “error” if a short segment of the posterior pericardium is not well visualized?

Error analysis, while novel, lacks the same degree of investigational rigor compared to other models and would be difficult to utilize in an ultrasound assessment. For these reasons, it will likely contribute little to our own assessment needs.

Literature Review: Ultrasound Skills Evaluation in the Literature

Many of the principles previously described have already been applied to ultrasound skills assessment in some capacity, but despite the recent interest in and expanded scope of ultrasound, relatively few studies have undertaken the challenge of developing an objective skills assessment model. Of those published works attempting to describe an objective assessment of ultrasound skills, few studies specifically assess ultrasound *imaging* performance. Instead, they typically consider the performance of an ultrasound-based procedure. While the paucity of ultrasound imaging assessment studies is disappointing we will be able to extrapolate some useful lessons from these related ultrasound-oriented papers.

Barrington et al. used a task-specific scoring system to assess the learning curve of performing an ultrasound-guided sciatic nerve block on a cadaver model⁷². Thirty candidates were taught and then subsequently asked to perform the procedure while they were recorded on video. Subsequent video analysis was completed to assess performance on a task-based scoring system. These scores amount to a hybrid of task-specific checklists and a global rating scale; the items were assessed on a four point quality-based scale using descriptions that were specific to ultrasound guided needle-based procedures, but were neither specific to this task nor generally applicable to all ultrasound procedures. Specifically, scores were assessed on the basis of needle-tip visualization during advancement through soft tissues, probe handling and steadiness, and visibility of the target during injection of medication. The goal of the test was to determine, via cumulative sum analysis (CUSUM), when participants reached technical proficiency. If we accept the theory that experience with the procedure contributes to expertise, then the steady improvement of scores on serial assessments could be said to demonstrate construct validity of their technical skills assessment. Agreement and reliability were not assessed in this study.

De Oliveira Filho et al. also studied the use of ultrasound in nerve blocks and used an animal model to perform a CUSUM of learning curves for peripheral nerve blocks⁷³. Using a task checklist that was more task-specific than Barrington et al.'s, participants were video recorded and subsequently scored on a four point scale on the number of needle punctures, time to complete the task, needle visualization during advancement, and overall quality of the performance. If we again accept the principle that experience contributes to expertise, then the progressive increase in performance scores with experience would represent construct validity demonstrated by this technical assessment. Evidence of improving expertise with experience is demonstrated by the task-completion times in the study, which decreased by 48% over six trials while quality scores increased by 59%.

Margarido et al. also assessed trainees learning a new task, this time asking eighteen anesthesiologists to learn the technique of ultrasound guided lumbar spine assessment. After providing reading material, the anesthesiologists participated in a ninety-minute education session to describe and demonstrate the technique as well as allow hands-on practice of the lumbar spine assessment on live human models. Subsequent attempts to perform the task were evaluated with regards to identification of insertion spaces, choice of optimal insertion point, and accuracy of measurement of distance from skin to ligamentum flavum. This represents a true task-specific assessment. As with Barrington and de Oliveira Filho et al., a CUSUM analysis was performed to assess the learning curve of the task. Unlike the previously described studies, participants did not measurably improve on serial examinations and the majority failed to demonstrate competence; while this does not contribute to an assessment of validity, it does show us that the measurements are not simply assessing experience, but some other variable (which we believe to

be competence). The use of live human models would contribute to the face and content validity of the study, and so the lack of construct validity demonstrated is regrettable.

Sites et al. once again assessed ultrasound learning curves by using ten very inexperienced junior residents to assess ultrasound guided breast cyst aspiration¹³³. A previously studied and validated bench model, composed of an olive embedded in a turkey breast, was used to assess performance¹³⁴. Participants were asked to perform a cyst aspiration on the model while under video recording. Outcome measures included time to complete the task, number of needle passes, number of (previously defined) errors, and image quality. Image quality was graded on a four point scale while the other variables were measured in number of occurrences. Over six trials, trial times decreased by 48% while overall quality scores improved by 57%, again giving us some evidence of objective improvement which may be a form of construct validity.

Unfortunately, all four of these studies lack comparison to an accepted standard of objective technical assessment and therefore are lacking in criterion validity. Because they lack even a subjective pass/fail assessment we cannot conclude criterion validity to be demonstrated with these assessments and construct validity is only weakly demonstrated. Face and content validity for all was high however. Overall this cohort of studies demonstrates merely an interest in ultrasound skills assessment; while the authors share a similar goal with us, their work lacks a standardized format and instead blends different modalities of assessment. This makes it difficult to ascertain which elements of their studies were strong and which were weak.

A small number of authors have followed strict protocols using previously described objective scoring and assessment techniques, allowing their work to be more quickly understood, analyzed, and compared to our own ambitions. For example, Nair et al. attempted to improve the

quality of assessment and feedback given to cardiology fellows who are required to be trained in transthoracic echocardiography³⁸. Their assessment consisted of an interpretive portion and a “scanning” (image acquisition) portion, which were independently assessed. Notably all twenty-two participants had previously (on traditional, subjective ITER assessments) been scored as “meeting” or “exceeding” expectations of training. Construct validity was established by comparing performance of fellows in ascending years of training and inter-rater scoring comparisons assessed reliability. Fellows were instructed to perform their “standard” examination of a patient presenting with “chest pain” for a “ventricular function assessment.” Live human models were used with one model having an expert-determined normal exam and one with a known wall-motion abnormality. Assessment utilized a task-specific checklist and a global rating scale with a pass mark established at 60% of both scales based on expert opinion of the rating systems. Fellow experience correlated significantly with scanning scores, but multiple segments of the scanning score (for example, quantitative Doppler) showed high failure rates. The inter-rater agreement with Cronbach alpha was 0.91 for scanning and overall 86% passed. This demonstrates that a rating scale and checklist have construct validity, content and face validity (as the exams were performed on real patients with real pathology), reliability, and criterion validity (as the high pass rate is reflected by the good subjective evaluations by the participants prior to the study). Despite some poorly performed segments of the exam, this study provides compelling evidence of both the validity of a task checklist and rating scale system and also their practical value, in that they highlight specific areas of deficiency not otherwise noted on the previously used subjective ITER assessments.

Sultan et al. developed a sixty-three point task specific checklist and an accompanying global rating scale to objectively assess performance of ultrasound-guided axillary brachial

plexus blockade¹³⁵. Participants were divided into novice, intermediate, and expert groups based on experience. Blocks were performed under expert supervision on live patients while being video recorded, and experts were instructed to take over the procedure for any of a defined number of safety concerns; the recordings were subsequently reviewed and scored by four experts in the technique. Overall, intra class correlation of the checklists calculated as Cronbach alpha values was 0.842 and of the rating scale was 0.795. Pairwise class analysis demonstrated significantly higher scores with experience for both the rating scale and checklist when comparing novice to intermediate, intermediate to expert, and novice to expert classes. This test demonstrates construct, face, content, and criterion validity (based on the live expert safety assessment) as well as reliability in objective ultrasound assessment.

Finally, Chin et al. validated the use of hand-motion analysis in the performance of ultrasound guided peripheral nerve blocks⁹². A junior cohort of residents was compared against a cohort of experienced experts in one arm of the study, while junior regional anesthesia fellows were compared against senior regional anesthesia fellows in a second arm. Each participant's procedure was performed on a human volunteer, video recorded, and scored by two experts. Assessments included hand motion analysis for number of movements, total path length travelled, and time to complete the procedure as well as assessment on video review via a seven-point global rating scale and a thirty-item task checklist. Comparing residents to attendings, all measures of hand motion reached statistical significance; comparing early versus late fellows, all measures of hand motion excepting left hand movements reached statistical significance. The checklist and global rating scale demonstrated Pearson correlation coefficients of 0.97 and 0.98 respectively; attending physicians scored significantly better on both tools than residents, and

late fellows scored significantly better than early fellows in both as well. Overall, this is an example of construct, content, face, and criterion validity as well as reliability.

These objective ultrasound skills assessments lack a purely image-acquisition derived assessment and instead measure competence with regards to the combination of imaging and interpretive or procedural skills. As noted in the introduction of this paper, an inability to generate high-quality ultrasound images will presumably make diagnosis or ultrasound-guided procedures difficult if not impossible. FAST is an excellent model to assess imaging skills and we hope to demonstrate this using some of the tools already proven to be valid in other ultrasound and non-ultrasound assessments. To do so, extensive modification of previously described rating scales and checklists will have to be undertaken to remove the irrelevant procedural components and create a pure imaging-based assessment. After reviewing the available literature, it is apparent that such a proposal is realistically feasible and has not yet been attempted by other authors.

Literature Review: Conclusions and Preliminary Assessment Tool Design

After considering the multiple different evidence-based assessment modalities at our disposal, we are left to determine the optimum means of conducting the assessment. In the previously described studies there have been a variety of test scenarios devised including live and real-time assessments, video-recorded assessments, combined live-and-recorded assessments, as well as modalities including performance on models, volunteers, and real patients. Content validity assesses the level of realism of the assessment technique, and thus the most valid assessments will always be those that involve the real application of a procedure on a real patient. While live “real world” assessments can be valid assessment methods¹³⁶, this is not always practical or safe, particularly when our target examination involves the investigation of hemodynamically unstable trauma patients.

While there is limited evidence specific to ultrasound investigations, comparisons of bench models versus live operative examinations do seem to demonstrate that there is a significant correlation between the scenarios. Datta et al. demonstrated in one comparison of saphenofemoral vein dissection, comparing live surgical procedures on a human volunteer to the use of an inanimate lab model, that there was strong correlation between both a global rating scale and task checklist on an OSATS-style assessment, arguing that bench models may therefore be an acceptable test medium as a proxy for live surgical performance¹³⁷. We can infer therefore that we need not evaluate FAST ultrasound performance on real trauma patients to validate our assessment technique so long as we have a realistic and high quality model. Given the safety of ultrasound, a human volunteer without traumatic hemodynamic instability is a reasonable model choice.

There is debate over the use of live assessments versus video recorded assessments; some authors suggests that live evaluation may introduce bias which is controlled by video recording¹³⁸, while others argue that video-recording adds a layer of anxiety or pressure to study participants that is unnatural and compromises performance⁶³. It must be noted that our specific test (FAST ultrasound) is indicated in the management of unstable major trauma victims which is itself a high-stress environment and therefore participants must be comfortable performing the exam without compromise regardless of personal stress levels. Further, the use of a video-recording based assessment is the most practical means of performing our test as it will allow reviewers to be thorough in reviewing each specific stage of the ultrasound exam in critical detail. The use of video recording in technical skill assessments has been previously demonstrated in regional anesthesia^{63,139,140}, resuscitation^{70,79}, laparoscopic surgery^{55,57,116}, ophthalmologic microsurgery⁶⁰, cardiac surgery⁵⁹ and arthroscopic surgery^{49,51}. Because a video of the performance may reveal the identity of the trainee and thus bias the outcome, care must be taken to make participants as anonymous as possible¹⁴¹. This can be achieved in a video recorded study by blinding the reviewers to the identity of participants by, for example, positioning the video camera to not capture the participants' faces or by having all participants wear a standard set of clothes so as to be indistinguishable from one another.

We are left with the following conclusions about the technical assessment of ultrasound image acquisition skills. Technical skills assessments must be valid, feasible, and reliable. Because live assessment using a trauma patient model is unfeasible and unsafe, assessments should be conducted in a simulated environment using as realistic a model as possible, preferably using random-ordered video recordings of anonymous participants to minimize overlooking subtle image generation nuances that may be missed on a live assessment. While global rating

scales, task-specific checklists, combined tests, hand motion analysis, simulation metrics and error assessment all appear to be valid and proven means of assessing technical skills, they do not all apply to our proposed target skill. No ultrasound simulation device has validated built-in metrics for skills assessment and thus this technology must be regarded as promising but unproven as yet. Error analysis is untested as an ultrasound assessment modality and may be best suited to errors that are gross and binary (for example, “knot is not square”) rather than ultrasound errors that may be more subjective (for example, “gain is not perfectly adjusted”). Thus, these two modalities may be removed from consideration.

Our proposal therefore, based on the evidence presented in this literature search, is to develop and validate an objective scoring assessment of FAST ultrasound performance using the following criteria:

1. A global rating scale assessing the quality of techniques universal to ultrasonography
2. A task-specific checklist assessing the critical components of FAST imaging
3. Electromagnetic hand motion assessment including measures of time, number of movements, and path length

Such a tool using these proven modalities would then be scrutinized and put to the test by comparison of a novice and expert cohort to validate its use as an objective assessment of ultrasound image acquisition skill.

Methods: Development of Scoring Systems

Development of an objective ultrasound imaging assessment tool which will make use of a global rating scale, a task checklist, and hand motion analysis required the development of novel scoring checklists and scales which have not been previously described in the literature. To develop these, first the process of FAST image acquisition was broken down into its constituent parts for independent evaluation. These parts are divided by anatomic regions and include:

1. Evaluation of the right upper quadrant (also known as Morrison's Pouch or the hepatorenal space)
2. Evaluation of the left upper quadrant (also known as the splenorenal space)
3. Evaluation of the pericardium and heart
4. Evaluation of the pelvis and bladder

Our Delphi technique recruited ten recognized experts in medical education, ultrasound, and FAST. The panel of experts was recruited in consultation with the Secretary and President of the Canadian Emergency Ultrasound Society. An initial proposal for three checklists was developed by the primary investigators of this study and circulated to the panel of experts via email with invitations to contribute constructive criticism and feedback. After compilation of the initial round of feedback the suggestions were considered and those items with multiple messages in support were amended per the panel's suggestions. A second round of opinion solicitation was carried out with the modified versions of the scoring systems and again, feedback was considered and incorporated into a final version of the three scoring checklists.

The dynamic imaging task checklist, titled for our purposes as the FAST Image Acquisition Checklist (Appendix A) is a binary scoring checklist granting candidates one point for each item successfully accomplished on the dynamic, real-time video recording of the

ultrasound device's output. This test best approximates "real-world" conditions of FAST interpretation as the typical FAST exam is viewed and interpreted live as it is being performed, without pausing to save individual images for later review. It includes subheadings for four anatomic regions as previously described with six critical items in each anatomic region. The maximum possible score is twenty-four, indicating that all six items in each anatomic region were successfully documented on the video-captured ultrasound outputs.

The static image task checklist, titled the FAST Representative Image Checklist (Appendix B) is also a binary scoring checklist where candidates receive one point for each item successfully accomplished on still images captured during the course of their sonographic examination. Modification with the Delphi technique resulted in a four anatomic subheading, sixteen-item scoring checklist. These critical items are asymmetrically distributed based on the expert consensus, with five points derived from the right upper quadrant views, five from the left upper quadrant views, two from the pelvis views, and four from the pericardial views. A total score of sixteen would again indicate the successful visualization of each item in each anatomic view as captured by the participant. While this test does not represent "real-world" conditions (as the FAST exam is interpreted from the real-time images generated by the examiner) its inclusion allows us to infer that trainees have an understanding of the structures they are seeing and are capable of manipulating the probe in such a way that they can successfully generate images matching specific requirements.

Our global rating scale, titled simply the Global Rating Scale of FAST Image Acquisition (Appendix C) is a five-point Likert scale assessing nine domains of ultrasound technique. Specific anchors were defined for scores of one, three, and five on the Likert scale within each domain. Notably, consensus opinion held that a score of four would represent the expected

performance of a sonographer who meets the expectations of safe practise within the domain and would be considered a “pass” while a score of five would represent an exemplary performance which exceeds expectations. One category, the “Overall” performance score, provides anchoring statements for all five levels on the Likert scale to discourage a purely subjective assessment and encourage reviewers to base scoring on witnessed technical or imaging deficiencies.

This combination of three items represents the human-scored component of our assessment of FAST image acquisition skill.

Methods: Hand Motion Analysis Hardware and Software

Hand motion analysis was completed using a trakSTAR 3D tracking device¹⁴². Raw data was analyzed using locally developed Motion Analysis and Recording Software v3.0 or MARS3 (University of Manitoba, Winnipeg, MB). The magnetic field generator component of the hand motion tracking hardware was always placed immediately adjacent to the greater trochanter of the volunteer patient's left femur.

The hand motion hardware and software tracks motion with six degrees of freedom. Software analysis converts data from raw scores of X, Y, and Z axis movements into meaningful measures such as number of movements, path length, time, acceleration, velocity, and jerk. To smooth velocity curves and eliminate background "noise," the software utilizes a low pass Gaussian filter with a width of 12 samples per standard deviation corresponding to a high frequency cutoff filter of 1.666 Hz when a sample rate of 240 Hz is used. The goal of this filter is to ensure that only higher amplitude meaningful slower movements are recognized by the device, rather than tremors. We used a velocity threshold for movement detection of 15 mm per second. These numbers were based on calibration tests and previously validated thresholds¹⁴³. For all variables, left and right hand characteristics were scored independently but analysis where described was performed using combined left-right data; for example, number of movements is scored for the left hand and right hand, but is also combined into a sum total where described.

Methods: Study Design and Enrolment

Our validation protocol required the recruitment of two cohorts of ultrasonographers, one novice and one expert, from our local medical school. Power calculations were carried out based on an effect size of 1.0 to 1.2 standard deviations when comparing groups of different levels of experience as suggested in the literature^{144,145}. Power to detect an effect size of 1.2 standard deviations with an alpha of 0.05 (two-tailed) and beta of 0.80 suggests that we required two cohorts with twelve subjects per arm.

All participants signed a consent form indicating their agreement to participate in this study including video, ultrasound image, and hand motion data recording in accordance with requirements of the University of Manitoba Health Research Ethic Board (Appendix D).

Participants for the study were recruited via email communication to targeted departments. Novices, defined as resident physicians with no formal training in FAST ultrasound and who do not themselves perform FAST ultrasonography, were recruited via email solicitation to the junior resident cohort of trainees in the Department of Surgery. Experts, defined as those with both credentialed formal training in FAST ultrasound *and* who self-reported their routine use FAST in their clinical practise were recruited via email solicitation to the departments of Surgery and Emergency Medicine.

Our study utilized a live human volunteer as an ultrasound subject. This twenty five year old male subject was utilized for all twenty four trials, had no previous surgeries and suffered no significant abdominal trauma or abdominal sepsis. Trials were conducted in our university affiliated simulation lab. The room was set up in advance with a consistent design for every study. The ultrasound device was connected to an Epiphan Lecture Recorder x2 (Epiphan

Systems Incorporated, Canada) video capture terminal which generated synchronized, side-by-side images of the live ultrasound output and the live camcorder output in split-screen fashion, resulting in a single time-calibrated video stream showing the participants' hand and body movements in real-time with the corresponding ultrasound image changes.

Participants completed a brief questionnaire regarding their personal experience with ultrasound in both educational and clinical use (Appendix E). All study participants were then asked to watch a five minute video published by the Hennepin County Medical Center¹⁴⁶ explaining the necessary views to obtain during FAST performance and the goals of imaging each abdominal and thoracic region of the standard FAST exam. Subsequently, participants were oriented to the ultrasound device at our disposal, a Sonosite M-Turbo¹⁴⁷. Instructions regarding the device included an explanation of how to manipulate gain, how to save still images, where to locate ultrasound gel, and how to change the depth of view. Participants were asked to conduct a complete FAST exam as demonstrated in the instructional video, and also were asked to capture one high-quality representative image from each anatomic region; if more than one still image was captured they were asked to indicate at the time of the exam which image they felt was the highest quality. After orientation to the device, participants could elect to watch the orientation video a second time but were not obligated to do so.

Immediately before starting their ultrasound exam, participants were asked to don gowns, gloves, and surgical masks. Hand motion tracking magnets were taped to the dorsal surface of their hands between the third and fourth metacarpals. Hand motion tracking and timing was initiated only at the moment that the ultrasound probe made first contact with our model's skin. Participants performed one complete FAST exam as per the instructions seen in the introductory video and captured four representative images during the examination. Timing, video, and hand

motion recording were terminated when the participant removed the probe for the final time from the volunteer's skin after completing the last anatomic region requiring imaging; participants were asked to announce when they had finished to ensure accuracy of the timing and recording.

Methods: Data Scoring and Video Review

After collecting all twenty four trials of data, two experts (who were not part of the expert cohort) were tasked with reviewing the recorded videos and scoring the performances according to the previously drafted scoring checklists. Reviewers were oriented to the evaluations forms and participated in a calibration session during which four sample videos were scored. Reviewers independently scored the four sample videos according to our scoring checklists, and subsequently re-watched the videos while discussing at the appropriate times during the videos which scores they had assigned to each anatomic region and why. Consensus was achieved through this discussion process as to the scoring thresholds that would be applied to the videos, particularly with regards to the non-binary GRS scoring system. Both reviewers expressed satisfaction at the end of this session, and were comfortable with the scoring models.

The twenty four trials were stripped of all identifying data and ordered according to a random number generator. Each reviewer was provided for each participant with copies of the scoring checklists, one video file containing the split-screen video of camcorder and ultrasound outputs, and four still images as selected by the participants. Reviewers were blind to the identity of participants and conducted their video and image review independently and without consultation between reviewers or any other researchers.

Methods: Statistical Analysis

Statistical analysis of all data was carried out using the Statistical Analysis Software (SAS) version 9.3 (SAS Institute, Cary NC) and/or R Statistical Library (R Core Team, 2013). In all cases we accepted an alpha of 0.05 by convention as indicative of significance. Substantial or better inter-rater agreement (kappa) results, with scores greater than 0.6, were felt to be significant.

Comparisons of means between the novice and expert cohorts were calculated using two-tailed Students T Tests. T Tests were used for comparison of the GRS scale, instead of using non-parametric tests, because our design was based on a true Likert scale with data intended to be evaluated as a sum with approximately equal graduations between individual scores within each domain.

Intra class correlation (ICC) results comparing inter-reviewer agreement was calculated using weighted Cohen's Kappa for the scores of anatomic regions where data fit a limited number of possible scores within a narrow range. The Shrout and Fleiss conventional Intraclass Correlation Coefficient was used in place of Cohen's Kappa for the summed scores of all anatomic regions because scores approximated a normal distribution.

Univariate logistic regression models were used to determine the predictive power of predictors with significant T Test results. Finally, the leave one out cross validation technique was applied to logistic regression models and the associated modeling error was calculated; this technique uses the logistic regression model to calculate estimated values for real data points in our dataset, and compares the error of the estimate versus the observed measures to estimate the accuracy of the model. For each logistic regression model sensitivity and specificity were calculated using

software (though no out-of-cohort dataset was recorded for comparison) and used to generate receiver operating curves (ROC) plotted as the sensitivity (y-axis) against 1-specificity (x-axis). The ROC curves were used to estimate the optimal discriminatory score via the conventional technique of selecting the inflection point of the resulting graphical curve. Finally, collinearity was assessed in multivariate modeling by estimating the variance inflation factor and assessing the standard error of the multivariate model.

Results: Introduction

In total twelve novice ultrasonographers and twelve expert ultrasonographers participated in the study (Table 1). Novice cohort participants ranged from post graduate years one to three and included trainees in general surgery, urology, and plastic surgery; the expert cohort consisted of eleven emergency medicine specialists and one general surgeon.

Regarding personal FAST experience, 83% of experts reported performing the FAST exam on at least a weekly basis compared to 0% of novices (Figure 1), and 100% of experts indicated that they performed an ultrasound exam of any type at least weekly compared to 0% of novices; 42% of experts reported at least daily use of point-of-care ultrasound imaging (Figure 2).

Table 1: Participant Demographics

Comparator	Novice Cohort N=12	Expert Cohort N=12	P value
Clinical Background			
Emergency Medicine	0%	91.7%	<0.0001
Surgery	100%	8.3%	/
Handedness			
Right	100%	83.3%	0.1522
Left	0%	8.3%	/
Ambidextrous	0%	8.3%	/
Ultrasound Experience			
Performed in Simulation	41.7%	100%	0.0007
Performed on Patient	58.3%	100%	0.0103
Formal US Training	8.3%	100%	<0.0001
FAST Experience			
Performed in Simulation	16.7%	100%	<0.0001
Performed on Patient	16.7%	100%	<0.0001
FAST Credentialed/Certified	0%	100%	<0.0001

Figure 1: Frequency of Ultrasound Usage Among Participants

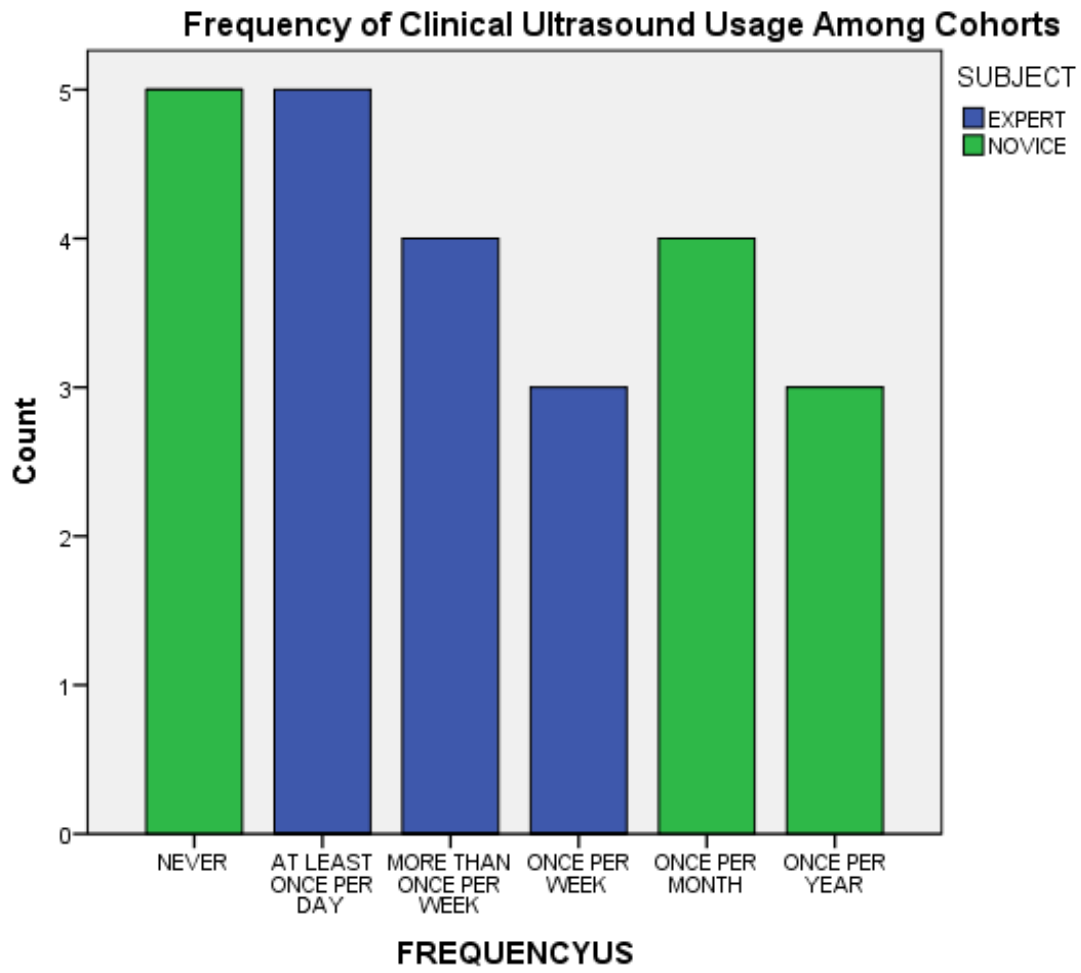
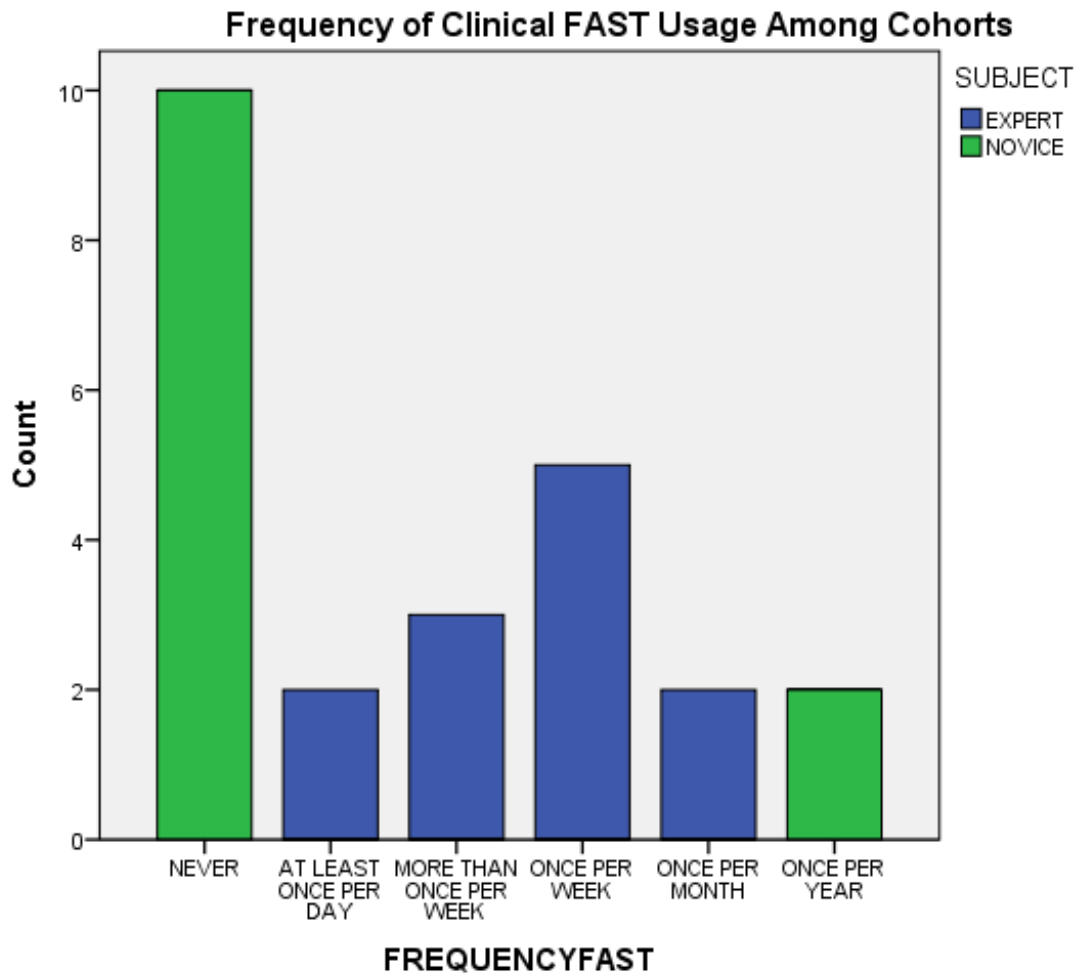


Figure 2: Frequency of FAST Usage Among Participants



Results: FAST Image Acquisition Checklist (Appendix A)

The FAST Image Acquisition Checklist results are summarized in Tables 2 through 6. Statistical analysis was conducted on the anatomic region total scores and the sum total; individual checklist item scores are demonstrated for reference though statistical analysis was conducted only in the regions noted.

Table 2: Hepatorenal Space Scores on FAST Image Acquisition Checklist

Hepatorenal Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Orients image with liver to left	0.75	1.00	/
Adjust image depth to below kidney	0.58	0.88	/
Adjusts gain appropriately	0.67	0.88	/
Visualizes interface between liver and kidney	0.75	0.96	/
Sweeps through entire liver/kidney	0.29	0.42	/
Visualizes caudal tip of liver	0.29	0.54	/
Sum of Scores for Region	3.33 (55.5%) (2.30 - 4.37)	4.67 (77.8%) (3.94 - 5.40)	0.0139

Table 3: Splenorenal Space Scores on FAST Image Acquisition Checklist

Splenorenal Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Orients image with spleen to left	0.71	1.00	/
Adjust image depth to below kidney	0.67	0.88	/
Adjusts gain appropriately	0.67	0.92	/
Visualizes interface between spleen and kidney	0.83	0.96	/
Sweeps through entire spleen/kidney	0.25	0.38	/
Visualizes space between diaphragm and spleen	0.29	0.50	/
Sum of Scores for Region	3.42 (57.0%) (2.35 - 4.48)	4.63 (77.2%) (3.87 - 5.38)	0.0283

Table 4: Pelvis Space Scores on FAST Image Acquisition Checklist

Pelvis Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Adjusts depth to 4cm below bladder	0.46	0.58	/
Adjusts gain appropriately	0.58	1.00	/
Visualizes bladder in longitudinal plane	0.71	0.71	/
Sweeps through bladder in longitudinal plane	0.25	0.38	/
Visualizes bladder in transverse plane	0.38	0.63	/
Sweeps through bladder in transverse plane	0.13	0.33	/
Sum of Scores for Region	2.46 (41.0%) (1.29 – 3.63)	3.63 (60.5%) (2.80 – 4.45)	0.0500

Table 5: Pericardium Space Scores on FAST Image Acquisition Checklist

Pericardium Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Orients image with ventricle to the right	0.63	0.92	/
Adjusts image depth to end deep to pericardium	0.17	0.88	/
Adjusts gain appropriately	0.46	0.88	/
Optimizes view using adjuncts	0.21	0.58	/
Visualizes both anterior and posterior pericardium	0.42	0.75	/
Sweeps through entire pericardium	0.00	0.33	/
Sum of Scores for Region	1.88 (31.4%) (0.86 – 2.89)	4.33 (72.2%) (3.61 – 5.05)	<0.0001

Table 6: Regional Score Summaries on FAST Image Acquisition Checklist

Anatomic Space Total Score	Mean Score, Novice (95% C.I.)	Mean Score, Expert (95% C.I.)	P Value
Hepatorenal Space	3.33	4.67	0.0139
Splenorenal Space	3.42	4.63	0.0283
Pelvis Space	2.46	3.63	0.0500
Pericardial Space	1.88	4.33	<0.0001
Sum of All Spaces	11.08 (46.2%) (7.97 – 14.20)	17.21 (71.7%) (15.00 – 19.41)	0.0005

Results: FAST Representative Image Checklist (Appendix B)

The FAST Representative Image Checklist results are summarized in Tables 7 through 11. Statistical analysis was conducted on the anatomic region total scores and the sum total; individual checklist item scores are demonstrated for reference though statistical analysis was conducted only in the regions noted.

Table 7: Hepatorenal Space Scores on FAST Representative Image Checklist

Hepatorenal Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Orients image with liver to left	0.63	1.00	/
Adjust image depth to below kidney	0.50	0.83	/
Adjusts gain appropriately	0.58	0.83	/
Visualizes interface between liver and kidney	0.38	0.83	/
Visualizes caudal tip of liver	0.04	0.29	/
Sum of Scores for Region	2.13 (42.6%) (1.25 – 3.00)	3.75 (75.0%) (3.13 – 4.37)	0.0008

Table 8: Splenorenal Space Scores on FAST Representative Image Checklist

Splenorenal Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Orients image with spleen to left	0.58	1.00	/
Adjust image depth to below kidney	0.63	0.71	/
Adjusts gain appropriately	0.71	0.83	/
Visualizes interface between spleen and kidney	0.50	0.88	/
Visualizes caudal tip of spleen	0.13	0.17	/
Sum of Scores for Region	2.54 (50.8%) (1.58 – 3.50)	3.58 (71.6%) (2.91 – 4.26)	0.0344

Table 9: Pelvis Space Scores on FAST Representative Image Checklist

Pelvis Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Adjust image depth to 4cm below bladder	0.50	0.58	/
Adjusts gain appropriately	0.71	1.00	/
Sum of Scores for Region	1.21 (60.5%) (0.69 – 1.72)	1.58 (79.0%) (1.22 – 1.95)	0.1454

Table 10: Pericardium Space Scores on FAST Representative Image Checklist

Pericardium Space Tasks	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Image oriented with apex to right	0.38	0.75	/
Adjusts depth to visualize superior pericardium	0.21	0.71	/
Adjusts gain appropriately	0.13	0.71	/
Both pericardium layers are visualized simultaneously	0.04	0.50	/
Sum of Scores for Region	0.75 (18.8%) (-0.02 – 1.65)	2.67 (66.8%) (2.03 – 3.30)	0.0002

Table 11: Regional Score Summaries on FAST Representative Image Checklist

Anatomic Space Total Score	Mean Score, Novice (95% C.I.)	Mean Score, Expert (95% C.I.)	P Value
Hepatorenal Space	2.13	3.75	0.0008
Splenorenal Space	2.54	3.58	0.0344
Pelvis Space	1.21	1.58	0.1454
Pericardial Space	0.75	2.67	0.0002
Sum of All Spaces	6.63 (41.4%) (4.50 – 8.75)	11.58 (72.4%) (10.08 – 13.09)	<0.0001

Results: Global Rating Scale of FAST Image Acquisition Checklist (Appendix C)

The Global Rating Scale of FAST Image Acquisition Checklist results are summarized in Table 12. Note that the “Autonomy” component of the rating scale was not assessed in this study but is a proposed item for inclusion in the score to be assessed in future studies.

Table 12: Outcome of Global Rating Scale of FAST Image Acquisition Scale as a Predictor of Expertise

Domain	Mean Scores, Novice (95% C.I.)	Mean Scores, Expert (95% C.I.)	P Value
Skin Contact	2.71 (2.15 – 3.27)	3.92 (3.52 – 4.31)	<0.0001
Image Adjustment	1.67 (0.96 – 2.37)	3.38 (2.88 – 3.87)	<0.0001
Initial Probe Placement	1.67 (1.13 – 2.21)	3.75 (3.37 – 4.13)	<0.0001
Image Sweeping	2.08 (1.48 – 2.68)	3.67 (3.24 – 4.09)	<0.0001
Sonographer Positioning	2.96 (2.42 – 3.50)	4.21 (3.83 – 4.59)	<0.0001
Time of Exam	2.13 (1.43 – 2.82)	3.29 (2.80 – 3.78)	0.0021
Flow of Procedure	3.79 (3.39 – 4.20)	4.25 (3.96 – 4.54)	0.0278
Autonomy †	N/A	N/A	N/A
Overall Score	1.42 (0.87 – 1.96)	3.33 (2.95 – 3.72)	<0.0001
Total Score	18.42 (15.14 – 21.70)	29.79 (27.47 – 32.11)	<0.0001

† Not assessed in this study but included in original scoring document

Results: Hand Motion Analysis

Hand motion details were documented in addition to the expert-scored checklists. One hand motion dataset was captured per participant, but one participant's electronic data file suffered corruption resulting in a total of twenty three data sets for twenty four participants.

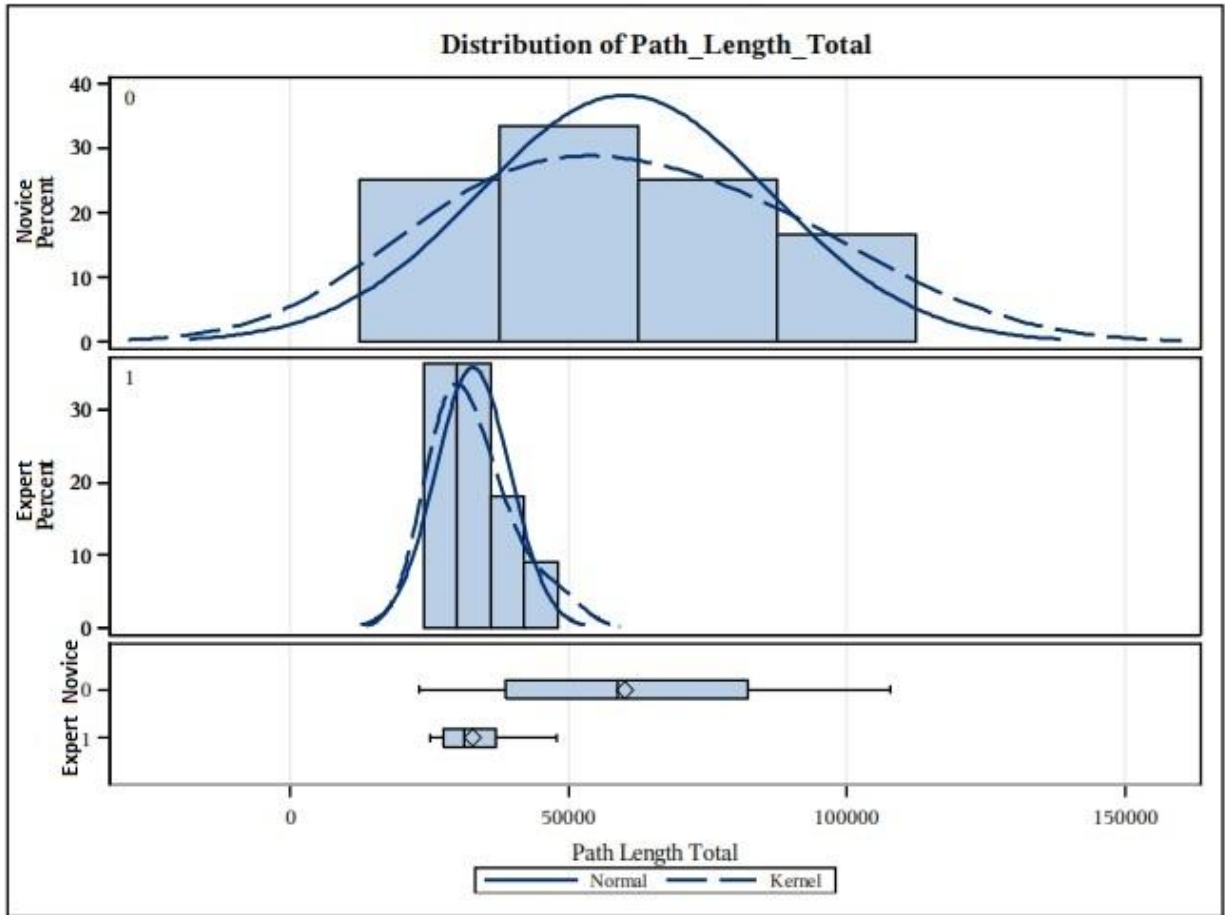
Length of exam (time) was assessed using the hand motion data; novices completed the exam in an average of 475.7 seconds compared to 345.9 seconds for experts without reaching statistical significance ($p=0.1219$).

Total path length of travel was measured for left hand, right hand, and total. These findings are summarized in Table 13. A significant difference in variance was noted between the two cohorts as demonstrated graphically in Figure 3.

Table 13: Path Length of Hand Motion by Novice and Expert FAST Performers

Measurement	Novice Distance (Std Dev)	Expert Distance (Std Dev)	P Value
Left Hand	28.010 m (13.388)	18.523 m (3.520)	0.0346
Right Hand	32.087 m (15.971)	14.254 m (4.232)	0.0026
Total (Sum Both Hands)	60.097 m (26.146)	32.777 m (6.680)	0.0041

Figure 3: Total Path Length of Hand Movement by Novice and Expert Cohorts (mm)



Number of movements was measured independently for left hand, right hand, and the sum total. Movements were calculated based on a change in position, a change in velocity, a change in acceleration, and/or a change in jerk (rate of change in acceleration). The number of discrete movements used by each cohort is demonstrated in Table 14.

Table 14: Number of Movements to Complete FAST Exam

Movement Criteria	Novice Movements Left Hand (Std Dev)	Expert Movements Left Hand (Std Dev)	P Value	Novice Movements Right Hand (Std Dev)	Expert Movements Right Hand (Std Dev)	P Value	Novice Total (Std Dev)	Expert Total (Std Dev)	P Value
Position	37.67 (29.11)	36.09 (21.20)	0.8844	35.67 (23.91)	28.36 (15.54)	0.3999	73.33 (48.09)	64.45 (41.68)	0.6173
Velocity	193.9 (111.00)	109.5 (38.66)	0.0269	258.5 (133.20)	153.5 (59.35)	0.0253	452.4 (238.00)	263.0 (83.70)	0.0216
Acceleration	329.4 (250.9)	373.8 (244.5)	0.6722	271.7 (191.1)	425.8 (325.4)	0.1761	601.1 (439.1)	799.6 (422.0)	0.3536
Jerk	359.3 (257.3)	516.2 (386.4)	0.2609	350.4 (234.0)	566.2 (433.9)	0.1479	709.8 (489.9)	1082.4 (817.4)	0.1946

Hand motion measurements of rate of change in multiple domains were also measured and are demonstrated in Table 15.

Table 15: Hand Motion Rate of Change Performance Observations

Measure	Novice	Expert	P-Value
Average Velocity	0.0643 m/s	0.0768 m/s	0.3719
Maximum Velocity Left	4.121 m/s	3.846 m/s	0.9150
Maximum Velocity Right	7.303 m/s	3.399 m/s	0.0421
Average Acceleration	0.480 m/s ²	1.018 m/s ²	0.0789
Maximum Acceleration Left	174.429 m/s ²	370.004 m/s ²	0.3060
Maximum Acceleration Right	514.387 m/s ²	211.448 m/s ²	0.0713
Average Jerk	127.298 m/s ³	562.975 m/s ³	0.0471
Maximum Jerk Left	1.579 x 10 ⁵ m/s ³	3.549 x 10 ⁵ m/s ³	0.2962
Maximum Jerk Right	4.865 x 10 ⁵ m/s ³	1.916 x 10 ⁵ m/s ³	0.0733
Movements per Second by Position, Total	0.1495	0.1861	0.0750
Movements per Second by Velocity, Total	0.9418	0.8051	0.0296
Movements per Second by Acceleration, Total	1.467	2.901	0.0769
Movements per Second by Jerk, Total	1.699	3.905	0.0579

Results: Inter-Rater Agreements

Weighted kappa statistics for the agreement of scores between reviewers within the FAST Image Acquisition Checklist demonstrated fair agreement within anatomic compartments and a substantial agreement of the total score (Table 16). The Intraclass Correlation Coefficient (ICC) for the total score of the checklist was 0.7951.

Table 16: Inter rater agreement on scoring of FAST Image Acquisition Checklist

Anatomic Region	Weighted Kappa Coefficient
Hepatorenal Space	0.3772
Splenorenal Space	0.3495
Pelvis Space	0.3548
Pericardium Space	0.4975
Total Score	0.7951*

*Shrout-Fleiss Intra Class Correlation Coefficient

Weighted kappa statistics for the Representative Image Checklist demonstrated moderate positive correlations between reviewer's scores on specific anatomic regions with a substantial agreement for the overall score (Table 17). The overall Shrout Fleiss Intra Class Correlation Coefficient was 0.7610.

Table 17: Inter rater agreement on scoring of FAST Representative Image Checklist

Anatomic Region	Kappa Score
Hepatorenal Space	0.4324
Splenorenal Space	0.4822
Pelvis Space	0.4882
Pericardium Space	0.4493
Total Score	0.7610*

*Shrout-Fleiss Intraclass Correlation Coefficient

Weighted kappa coefficients for each domain of the Global Rating Scale fell in a range from slight to substantial agreement with a total score representing a substantial agreement (Table 18). The overall Shrout Fleiss Intra Class Correlation was 0.6066 Note that, as

previously described, the “Autonomy” domain was included in our proposed algorithm for future use as an assessment tool of new trainees but is not assessed in the present study. Modified data was considered as well, with scores of “5” or “exceptional” performance re-coded to scores of “4” or “passing” performance to reassess agreement on a 4 point Likert scale.

Table 18: Inter rater agreement on scoring Global Rating Scale of FAST Image Acquisition

Domain	Weighted Kappa	Weighted Kappa, Modified Data
Skin Contact	0.1365	0.162
Image Adjustment	0.6700	0.831
Initial Probe Placement	0.3404	0.622
Image Sweeping	0.3158	0.508
Positioning and Probe Handling	0.2258	0.553
Time to Complete	0.1874	0.338
Flow of Procedure	0.0000	0.0000
Overall Performance	0.5152	0.725
Autonomy †	N/A	N/A
Total GRS Performance	0.6066*	0.860*

† Not assessed in this study but included in original scoring document

* Shrout Fleiss Intra Class Correlation Coefficient

Results: Logistic Regression Analysis

Data reaching statistical significance on the previously described analyses were considered for logistic regression analysis in both univariate and multivariate modeling (Table 19). In the instance of hand motion data modeling the previously described missing data set for one of twenty four participants was handled via casewise deletion.

Table 19: Univariate Logistic Regression Equations

Variable	Intercept (Standard Error)	Coefficient (Standard Error)	Odds Ratio	P value (95% Confidence Interval of Odds Ratio)	Area Under Receiver Operating Curve
Image Acquisition Checklist Total Score	-6.988 (2.745)	0.493 (0.186)	1.637	0.0080 (1.137, 2.355)	0.8988
Representative Image Checklist Total Score	-9.579 (4.160)	1.042 (0.428)	2.806	0.0158 (1.214, 6.488)	0.9226
Global Rating Scale Total Score	-17.026 (8.00)	0.6922 (0.314)	1.998	0.0274 (1.080, 3.696)	0.9762
Path Length Total	4.865 (2.032)	-0.00012 (0.000053)	1.000	0.0275 (1.000, 1.000)	0.8269
Total Movements by Velocity	3.915 (1.760)	-0.0118 (0.00549)	0.988	0.0323 (0.978, 0.999)	0.8205

The sensitivity and specificity for the GRS and Image Acquisition performance were calculated via an automated process performed by the R statistical software. The “Best” performance was taken to be the score yielding sensitivity and specificity figures closest to the

upper left corner of the ROC curve. The “Standard” performance was taken to be the score yielding a predictive value of 50%. The characteristics of the two performance inflection points are described in Table 20 while Receiver Operating Curves for our univariate analyses are demonstrated in Figures 4 through 8.

Table 20: ROC Inflection Points for Global Rating Scale and Image Acquisition Checklist

Scoring System	Probability of Expertise	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value
Global Rating Scale	/	/	/	/	/
Best Performance	37%	1.0000	0.9167	0.9333	1.0000
Standard Performance	50%	0.9286	0.9167	0.9286	0.9167
Image Acquisition Checklist	/	/	/	/	/
Best Performance	57%	0.7857	0.8333	0.8462	0.7692
Standard Performance	50%	0.8571	0.7500	0.8000	0.8182

The probability of expert classification (based on our logistic regression equations) allows us to interpret the above data into numeric values attainable on the scoring checklists, given the known logistic regression property that:

$$\text{Probability} = e^{(\text{error} + \text{coefficient} * \text{score})} / [1 + e^{(\text{error} + \text{coefficient} * \text{score})}]$$

From the above, GRS performance achieving a score of 23 represents a 24.9% probability of expertise, a score of 24 represents a 39.8% probability of expertise, and a score of 25 represents a 56.9% probability of expertise. With fractional scores being impossible on our

scoring system, we find that the GRS Best Performance score (the score with a probability greater than or equal to 37% probability) is 24 and the GRS Standard Performance score is 25.

Using the same technique, we know that Image Acquisition Checklist scores of 14 represent a 47.9% probability of expertise and scores of 15 represent a 60.0% probability of expertise. Therefore the Image Acquisition Checklist Best and Standard Performance scores are both equal to 15.

Figure 4: Image Acquisition Checklist Model Receiver Operating Curve

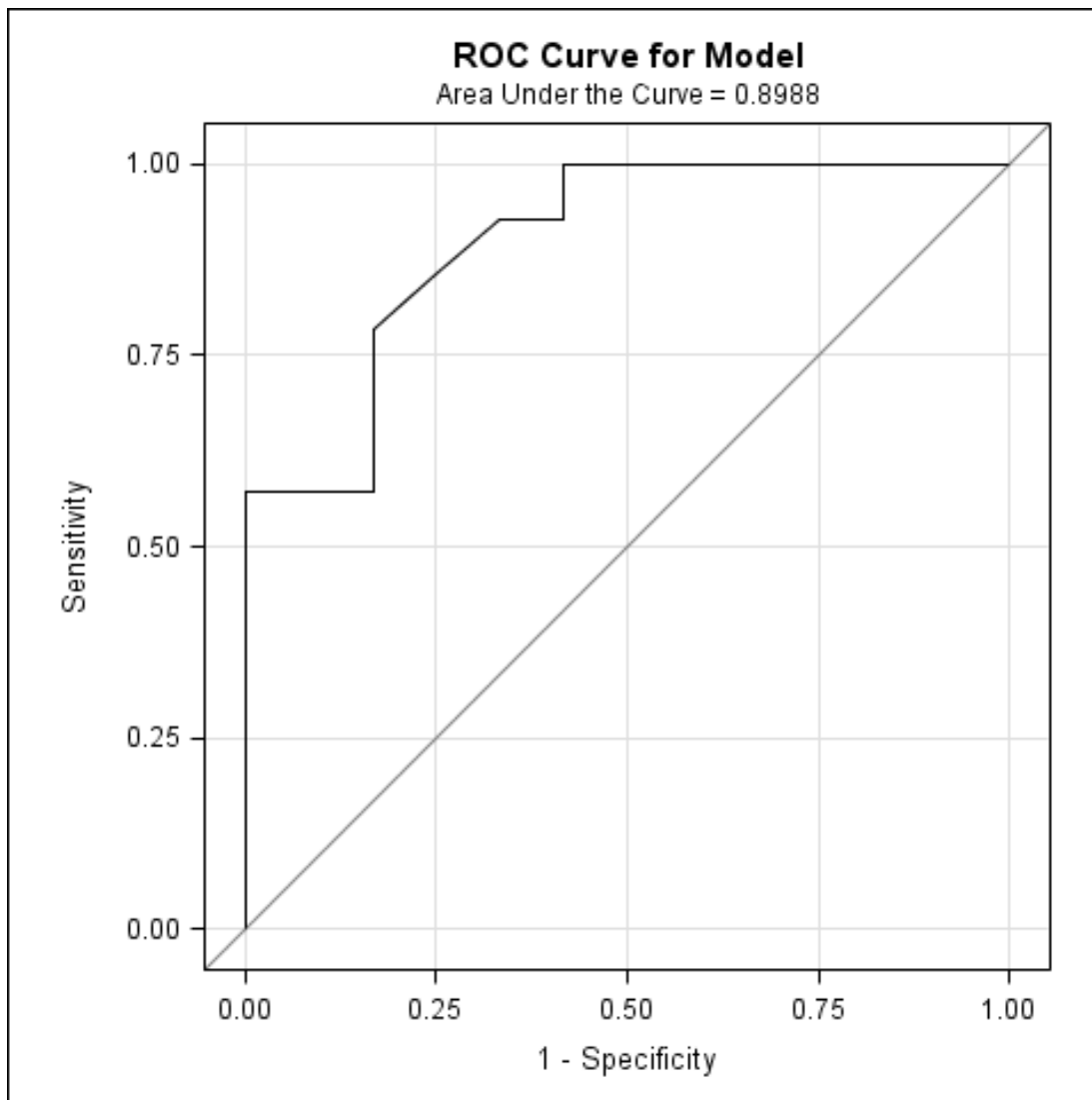


Figure 5: Representative Image Checklist Model Receiver Operating Curve

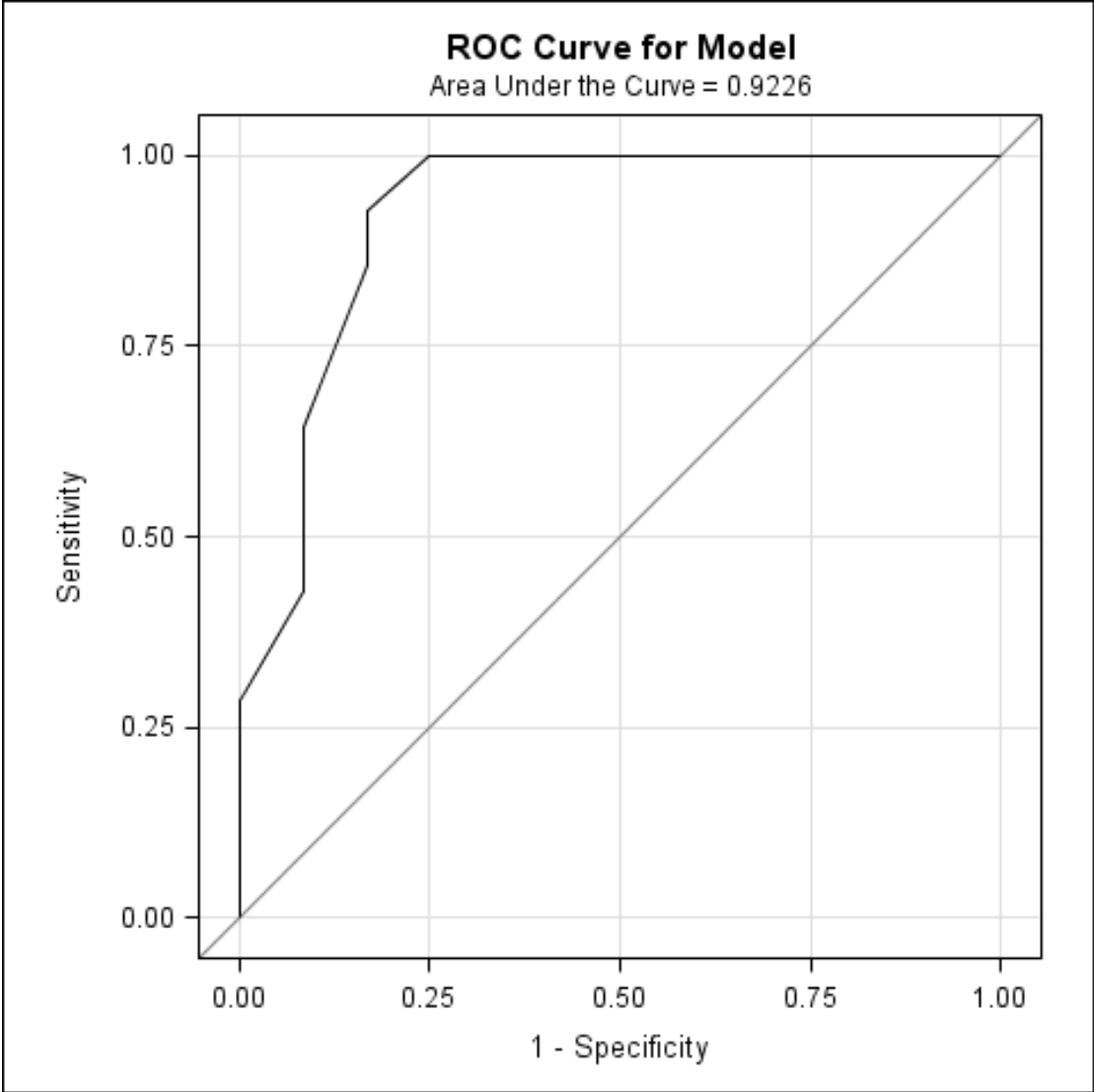


Figure 6: Global Rating Scale Model Receiver Operating Curve

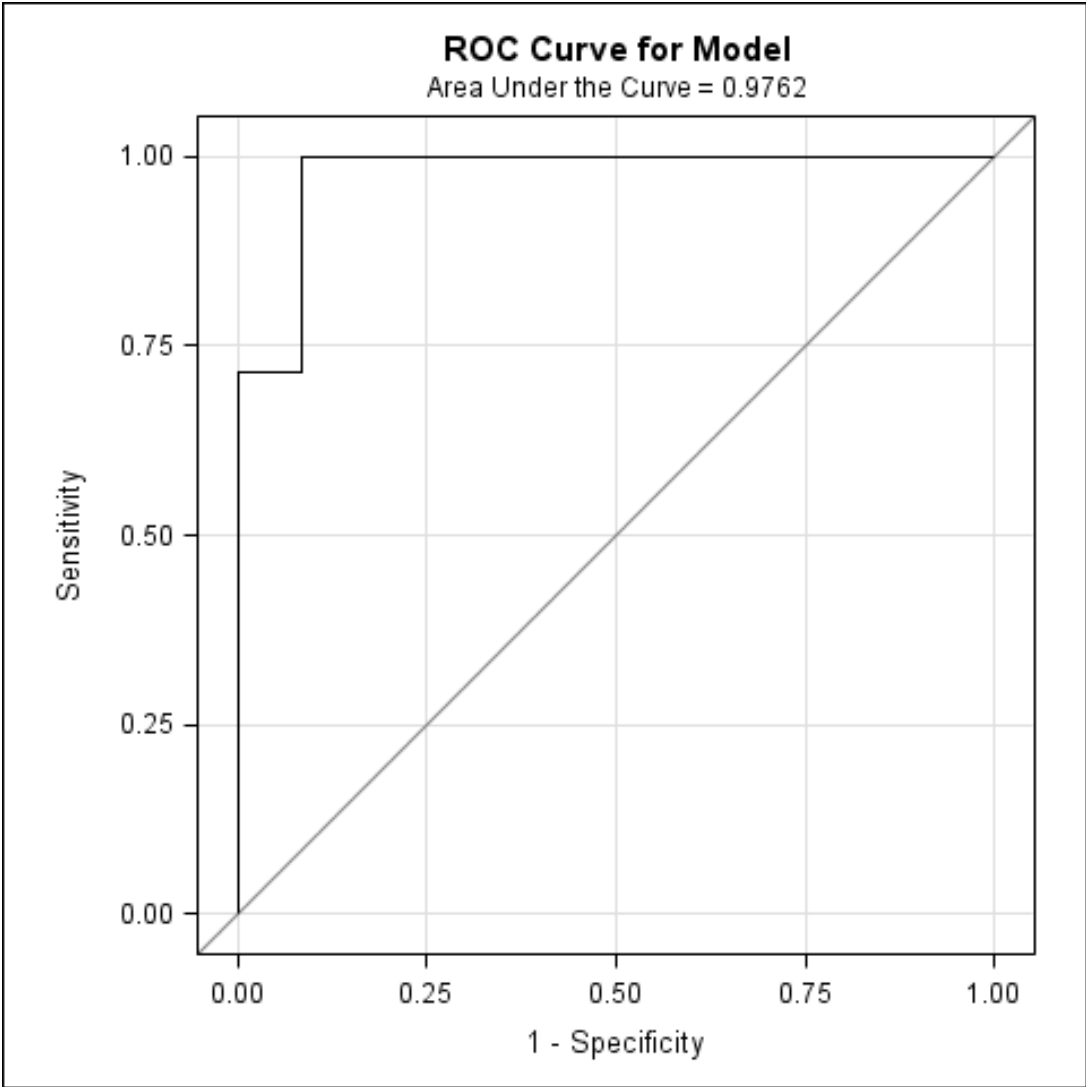


Figure 7: Total Path Length Model Receiver Operating Curve

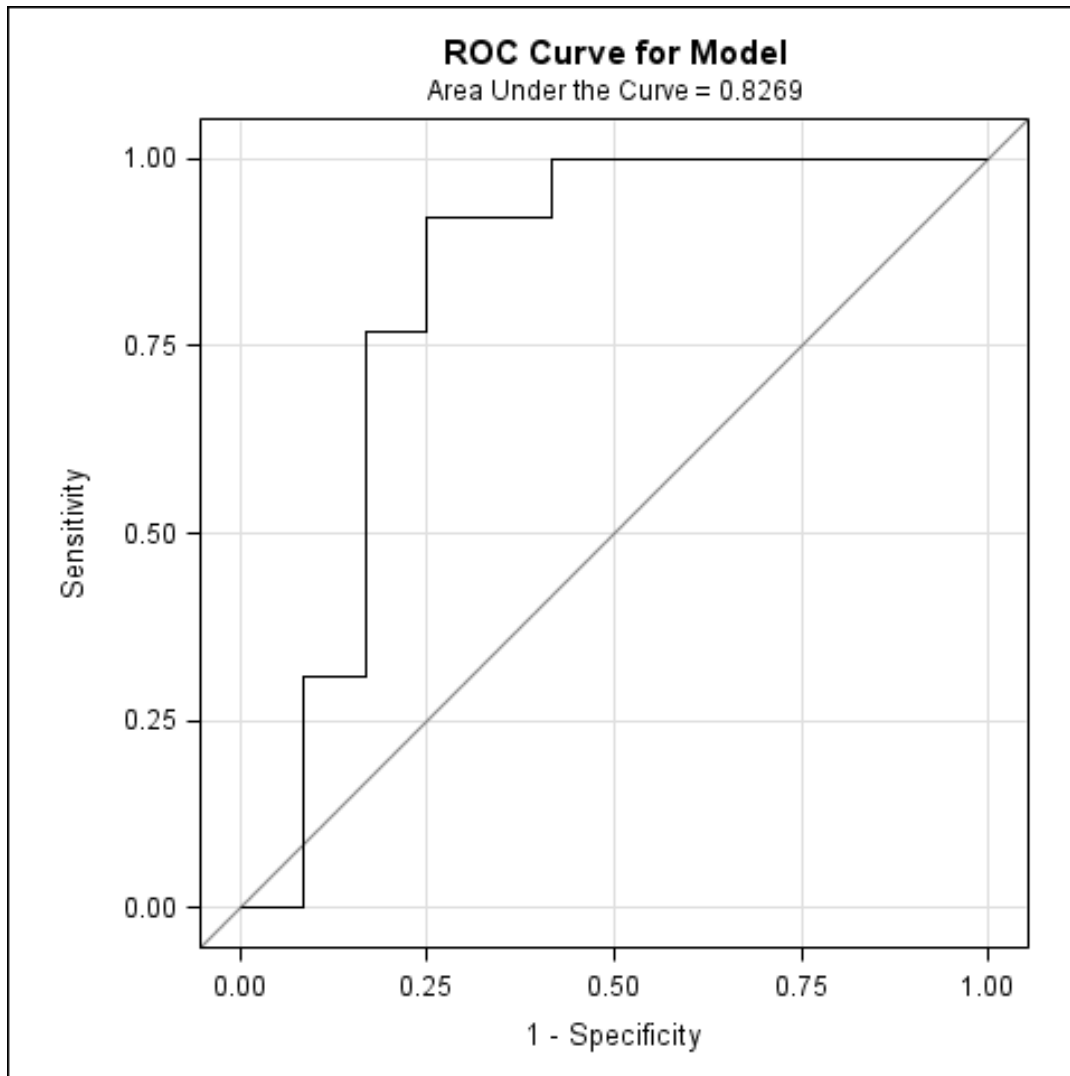
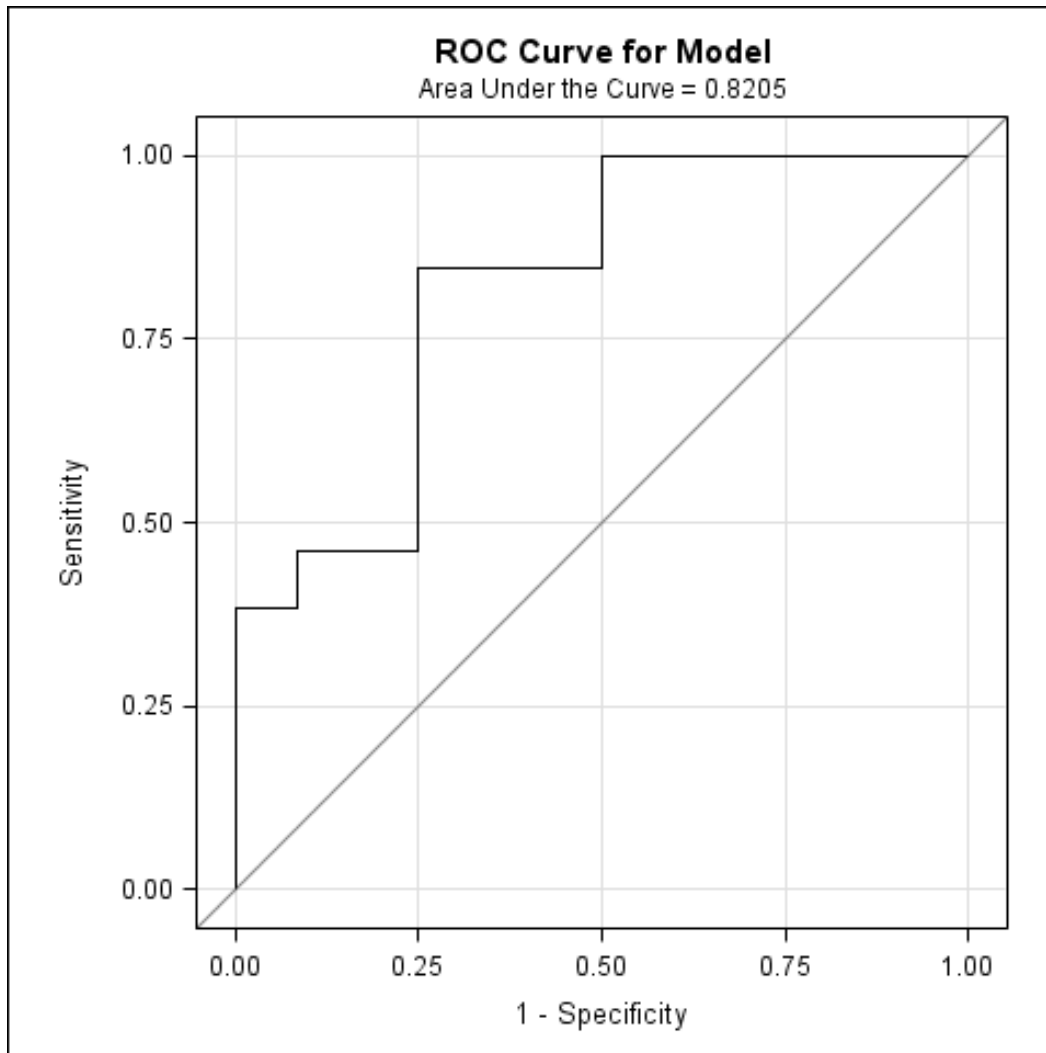


Figure 8: Total Hand Movements Model Receiver Operating Curve



Logistic regression models were validated by leave-one-out cross validation methodology which deletes a single data point from each category and attempts to predict the missing value using the fitted model¹⁴⁸ (Table 21). The adjustment in prediction error is an automated R feature which accounts for a bias introduced by modeling using fewer groups than there are observations¹⁴⁸.

Table 21: Leave-One-Out Validation Model Prediction Errors

Model	Prediction Error	Adjusted Prediction Error
Image Acquisition Checklist	0.2308	0.2308
Representative Image Checklist	0.1154	0.1139
Global Rating Scale	0.07692	0.08136
Total Path Length	0.2000	0.2016
Total Movements	0.2800	0.2800

Multivariate modeling was attempted using all five significant univariate predictors but suffered substantial variance inflation due to collinearity (Table 22).

Table 22: Multivariate Logistic Regression Model of All Significant Univariate Scores

Variable	Estimated Coefficient	Standard Error	Odds Ratio	Variance Inflation Factor
Intercept	-1.43	30131.20	n/a	n/a
Image Acquisition Checklist	-84.08	181031.66	3.052	5.933
Representative Image Checklist	76.20	117224.88	1.24	2.783
Global Rating Scale	120.33	189783.91	1.814	5.484
Total Path Length	13.02	104947.87	4.5135×10^5	6.991
Total Movements	24.57	135847.90	4.684×10^{10}	6.349

Multivariate modeling was attempted using only the Global Rating Scale and Image Acquisition Checklist predictors in the model to reduce collinearity (Table 23).

Table 23: Multivariate Logistic Regression Model of GRS and Image Acquisition Performance

Variable	Estimated Coefficient	Standard Error	Odds Ratio	Variance Inflation Factor
Intercept	-16.637	7.962	n/a	n/a
Image Acquisition Checklist	0.036	0.268	1.037	1.383
Global Rating Scale	0.654	0.365	1.923	1.383

Discussion: Proposal, Design, and Methods

We endeavoured to develop an objective tool to assess and score the quality of ultrasound image acquisition using a model built around the Focused Assessment with Sonography for Trauma (FAST). Because no previously validated model of objective assessment exists within the literature for FAST, development of the de novo model required us to dissect the FAST exam into its constituent components based around anatomic regions.

We worked under the belief that any objective assessment of a FAST ultrasound exam should include specific tasks and assessments of each anatomic region that must be satisfactorily imaged to demonstrate a passing performance. As the anatomic landmarks in each region are unique to that portion of the exam (for example, the liver tip need only be imaged in the right upper quadrant, not in every compartment) each region required the development of a unique scoring checklist related to the anatomic findings of that region. The successful imaging of these predetermined key views is best suited to analysis with a task checklist tool.

After dividing the FAST exam into anatomic quadrants, the method of imaging must be assessed. FAST ultrasound images are typically assessed by a clinician reviewing the live-streaming images directly on the sonograph display in real-time. In order for our test to demonstrate content validity the test scenario should include an evaluation of the real-time dynamic images captured by participants. Many ultrasound skill assessments however ask trainees to make inferences from static images; for our test to demonstrate criterion validity, we should also assess a candidate's ability to produce high quality representative images. This demonstrates an individual's understanding of the test, anatomy, and ultrasound technique as it forces the candidate to use this knowledge in a conscious effort to obtain the requisite views.

Theoretically, a novice sonographer could place the ultrasound probe on an abdomen and indiscriminately move it about in such a way that every conceivable component of an anatomic region is imaged; while such a manoeuvre would score points in a binary assessment of having imaged (versus not-imaged) critical structures it could hardly be said to be representative of ultrasound competence. Asking subjects to capture specific still images with defined criteria however does allow us to infer an understanding and a technical capability to capture the images requested, and it stands therefore that including a separate static image assessment is necessary. In both cases a task checklist can be used to document the success or failure of achieving requisite views, and so both static and dynamic ultrasound images required assessment with unique task checklists.

As an OSATS-style proposal, it stands that a global rating scale must also be a component of the exam and include general domains that evaluate a skill set across all ultrasound imaging techniques (that is, not specific to FAST). These skills should assess a trainee's familiarity with the ultrasound device, his or her understanding of how to use the equipment, how to maximize image quality, and how to perform the exam in a well-planned and progressive manner. The skills assessed should be well represented in a FAST ultrasound image capture but not specific to FAST, potentially allowing the same GRS device to be utilized in assessing other ultrasound-based investigations in future studies. The GRS component mainly evaluates technique and style components of the exam such as gain adjustment, continuity of skin contact and smoothness of movements.

In summary, three individual scoring systems were proposed and developed:

1. A global rating scale assessing ultrasonography skills in a task non-specific format
2. A task-specific checklist scoring specific critical items viewed on dynamic (real-time) imaging

3. A task-specific checklist scoring specific critical items viewed on static imaging

Development of these three scoring checklists was carried out through a modified Delphi technique, which is a widely accepted means of achieving consensus^{149,150}. Many objective assessments previously described in the literature have developed their scoring systems via a Delphi technique with good effect^{28,51,78,79,151}. Ultimately we developed the scoring systems summarized in Appendices A, B, and C.

Discussion: Data, Results, and Comparison to the Literature

The results of our pre-participation survey of both novice and expert cohorts supports our selection process and is in keeping with the “expert” and “novice” labels. Overall, experts made more frequent use of both FAST and other ultrasound investigations compared to the novice cohort who seemingly lack experience with this technology.

Comparing the mean scores of each checklist found that the Global Rating Scale, Image Acquisition Checklist, and Representative Image Checklists demonstrated statistically significant differences between the expert and novice cohorts’ performances (Tables 12, 6, and 11). Hand motion analysis was equally valuable, demonstrating significant differences between cohorts for both the total path length travelled and the number of movements (defined by a velocity gateway) (Table 13 and 14).

Agreement between reviewers in their independent assessments of performance demonstrated substantial agreement, with weighted kappas of total scores ranging from 0.6066 to 0.7951 for the original scoring system and from 0.7610 to 0.860 for the non-exceptional GRS scoring system. Comparison of agreement was conducted primarily using the weighted kappa statistic which also accounts for the potential of similar scores to be reached by chance alone. The kappa statistic ranges from -1 (perfect disagreement) to 1 (perfect agreement) with a score of 0 representing the expected result due to chance. The strength of agreements has been previously summarized (Table 24).

Table 24: Kappa correlation and clinical interpretation from Viera et al.¹⁵²

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

Overall, our findings are in keeping with the only other similar studies published in the literature to date, the works of Nair et al., Chin et al., and Sultan et al.^{38,92,135} These studies utilized an OSATs style assessment to assess the technical capabilities of novice and expert cohorts in the ultrasound-guided insertion of nerve blocks; Chin et al. included the notable additional measure of hand motion analysis using the ICSAD device. Unfortunately, the Nair study's primary focus is not the validation of their scoring algorithm; neither scores for the task checklist or global rating scale are provided, and the scoring system is not provided as an appendix. The Nair et al. study furthermore did not compare novices to experts, but compared two intermediate performing groups (second versus third year fellows). We are unable to include the Nair et al. study in our comparison due to insufficient published results and inappropriate cohorts for comparison. Our findings, namely that task checklists, global rating scales, and hand motion analysis are each independently valid methods of assessing technical skills proficiency in ultrasound imaging, are in keeping with the results of both the Chin and Sultan studies as summarized in Table 25.

Table 25: Comparison of Task Checklist Findings Between Similar Ultrasound Skills Assessment Studies

Assessment Tool	Score As a Percent of Total	Ratio of Expert to Novice Score	Inter Rater Agreement of Findings
Representative Image Checklist, Experts	72.4%	1.75	0.7610
Representative Image Checklist, Novices	41.4%		
Image Acquisition Checklist, Experts	71.7%	1.55	0.7951
Image Acquisition Checklist, Novices	46.2%		
Sultan et al. Task Checklist, Experts	87.1%	1.98	0.842
Sultan et al. Task Checklist, Intermediate	71.4%		
Sultan et al. Task Checklist, Novices	43.9%		
Chin et al. Task Checklist, Experts	93.1%	1.39	0.97
Chin et al. Task Checklist, “Late” Fellows	91.0%		
Chin et al. Task Checklist, “Early” Fellows	71.50%		
Chin et al. Task Checklist, Novices	66.8%		

92,135

Reviewing our task checklists compared to the literature findings we do see that our trial appears to agree with other reported studies. Both task checklists developed for our scoring protocol achieved expert to novice ratios that fall between the reported values in Chin and Sultan

et al. Our study seems to agree on a quantitative basis with the difference between expert and novice ultrasound performance on task-checklist assessment.

Considering the same comparison utilizing the global rating scale we find again that our results are comparable to the reported literature values. Again using Chin and Sultan et al. as our best comparable studies, we find global rating scale scores as follows outlined in Table 26:

Table 26: 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
GRS of FAST Image Acquisition, Experts	74.5%	1.62	0.607
GRS of FAST Image Acquisition, Novices	46.1%		
Modified GRS of FAST Image Acquisition, Experts	88.2%	1.54	0.860
Modified GRS of FAST Image Acquisition, Novices	57.2%		
Sultan et al. GRS, Experts	78.6%	1.97	0.795
Sultan et al. GRS, Intermediate	52.6%		
Sultan et al. GRS, Novices	39.8%		
Chin et al. GRS, Experts	94.9%	1.67	0.98
Chin et al. GRS, "Late" Fellows	92.1%		
Chin et al. GRS, "Early Fellows	65.7%		

Chin et al. GRS, Novices	56.9%		
-----------------------------	-------	--	--

The findings in comparing our global rating scale scores to the published results of other OSATS-style ultrasound assessments are generally in keeping with the previously identified patterns for our task specific checklist. While our original data set demonstrates the lowest overall scores, expert-novice ratios and inter-rater agreements, our standard and modified data is similar to the studies of Sultan and Chin et al. There are several possible explanations for these lower scores that relate to the choice of assessment technique (FAST versus peripheral nerve blockade) and the scoring system proposed (contents of the global rating scale used).

FAST certification in Canada typically is awarded on the basis of fifty observed assessments over two days and is frequently offered to junior residents. The FAST exam is performed frequently by certified individuals, as demonstrated in our pre-participation survey, and this too may lead to individuals developing unique strategies and short cuts which undermine the original principles of examination taught during certification training. This may lead to lower scores by experts, contributing to a lower expert-novice ratio. Peripheral nerve blockade, by contrast, requires longer periods of training and is performed less frequently by practitioners. We might hypothesize that prolonged training and infrequency of use contribute to an increased rate of adherence to training principles and decreased individualization of techniques. This should ultimately contribute to the greater variance between experts and novices demonstrated in Chin and Sultan et al., as well as the higher raw scores reported.

The unmodified GRS inter-rater agreement in our study was also the lowest of the three studies. Sultan and Chin utilized GRS scales without components of imaging assessment and

instead evaluate only technical steps of nerve blockade by anesthetic injection, while our own GRS measured only imaging technique domains. The difference in domains of the global rating scale assessments may thus be related to overall greater reviewer consensus on the procedural aspects of the study, while reviewer consensus on pure ultrasound imaging domains is not well assessed in the Sultan or Chin papers. Regardless of being ranked lowest, our results still fall in the “substantial” category of agreement.

We considered modifying our GRS data (by eliminating the “exceptional” category and reclassifying those scores to the “pass” category) in order to determine whether our agreement between reviewers was negatively biased by having two competing passing categories. While not quantified in the data, reviewers independently commented upon submission of their scores that it was often difficult to differentiate between the scores of “4” and scores of “5” on the video review. Because the agreement analysis equates to a lesser score for any disagreement on scores – even when both scores represent a passing grade – we attempted to quantify the level of disparity by categorizing all passing scores as one value. The difference in techniques is demonstrated in Table 18.

Moving forward however, we believe that the 5 point Likert scale is still the superior format as there are multiple advantages gained with the ability to appropriately score an exceptional performance. There are a multitude of potential uses to the objective identification of “super-experts” who surpass the minimum criteria for “expert” status. This category could be used to compare known or suspected super-experts against our current experts (for example, comparing radiologists against point-of-care ultrasonographers) in future studies, or could be used to identify which among the known experts would be potential candidates for more advanced training or even instructor status. Eliminating the “exceptional” category also raises

questions about the utility of a Likert scale where participants are expected to score maximally in order to pass; why not collapse the three failing grades into a single score as well and have a binary pass/fail protocol? Part of the value of a multi-point Likert scale stems from the ability to chart progression and create rankings; just as a failing trainee may demonstrate improvement, a successful trainee may also demonstrate improvements, and there is educational value in quantifying both findings. The value of our modified GRS scale, therefore, is purely to quantify the level of rater-disagreement attributable to passing-score-discrepancies, but the original *unmodified* GRS is the model that we would propose to use in future trials.

One element of our scoring system worthy of consideration is the evaluation of gain. Within the Image Acquisition and Representative Image checklists, gain is assessed independently in each of four anatomic domains. It was suggested during our scoring system development that this may introduce an element of bias into our assessment. If individuals set gain initially and do not adjust the gain setting throughout their performance, this may theoretically lead to an artificial inflation of the points awarded for the performance. A good initial gain setting would result in four points for one adjustment, and a poor initial gain setting would result in zero points for one adjustment. Alternatively, if participants were found to adjust gain frequently, this element of potential bias is overcome.

To address this issue, we examined whether or not participants were adjusting the gain throughout their performance in order to answer the question of what to do about potential gain bias. To do this we examined the distribution of gain scores awarded to the two cohorts (Figure 9, 10).

Figure 9: Distribution of Total Gain Scores Within the Representative Image Checklist

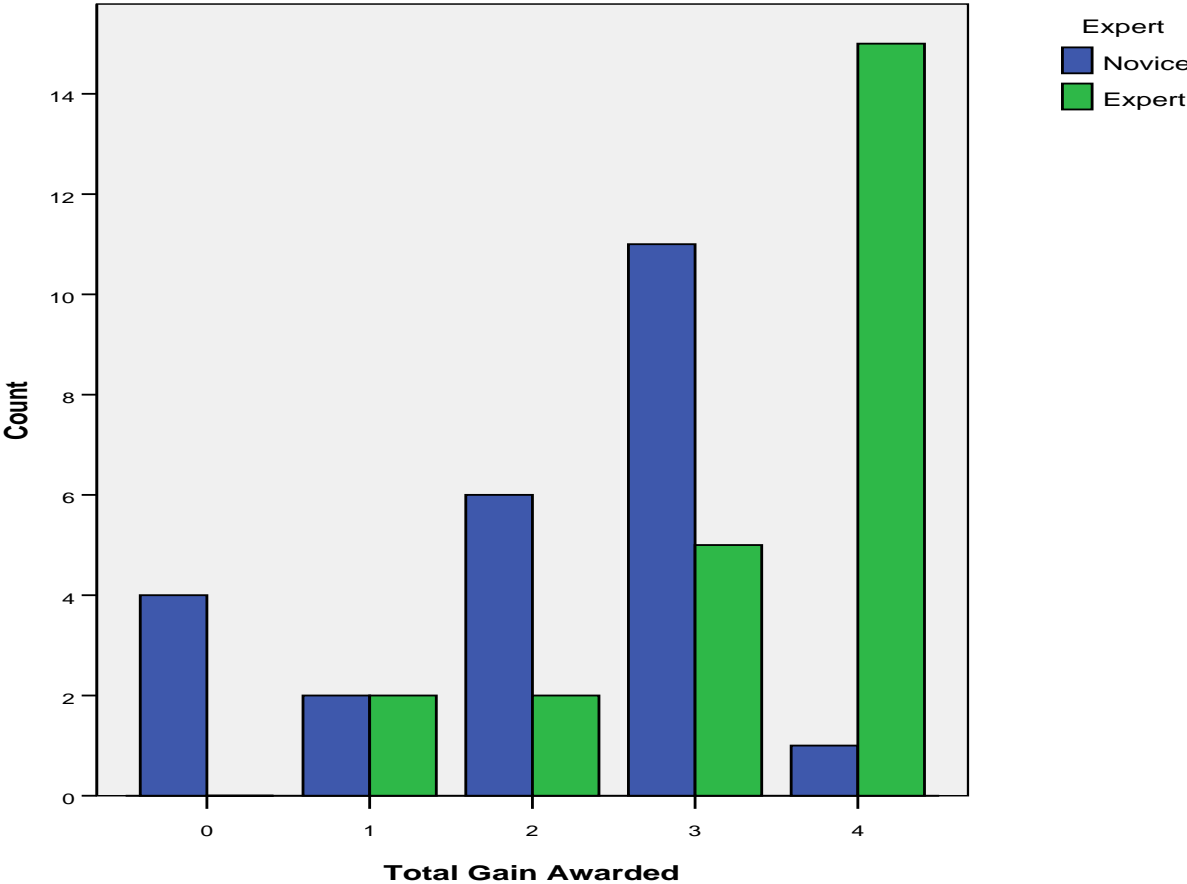
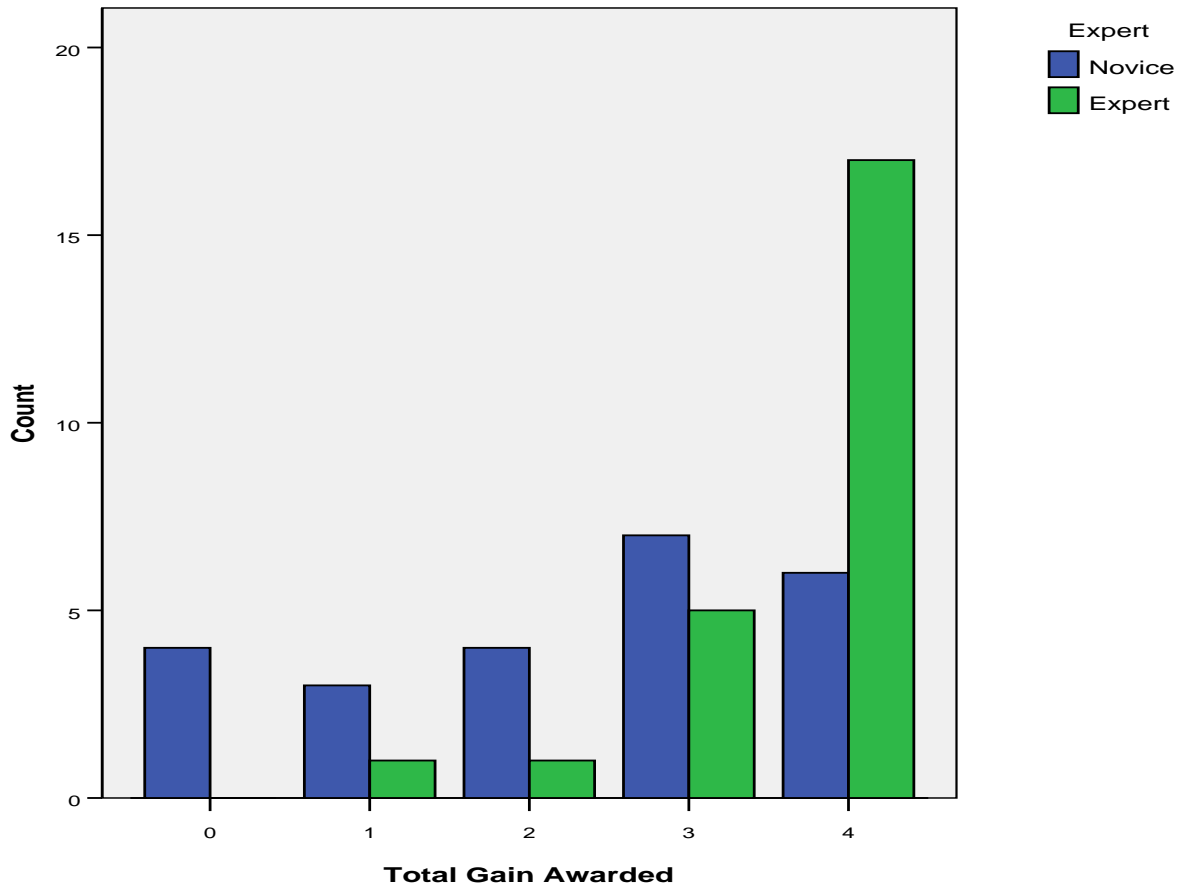


Figure 10: Distribution of Total Gain Scores Within the Image Acquisition Checklist



We also calculated the intraclass correlation of gain scores awarded, divided by reviewer and expert status (Table 27). Intraclass correlation measures the degree of correlation between repeated measures of gain in each trial.

Table 27: Intraclass Correlation of Gain Scores Divided by Class and Reviewer

Reviewer	Novice	Expert
Reviewer 0	0.3593	0.4915
Reviewer 1	0.1978	0.3179

If participants set the gain a single time and did not change it, we would expect the correlation between individual region gain scores and total gain scores to approach 1.00. This is not the trend demonstrated in our data set, with actual results demonstrating only mild to moderate correlations. On our graphical representation of gain points awarded, we would also expect to see dense clusters of scores of either zero or four points without intermediate results, a trend which is also not demonstrated in our data. Overall, the findings are not in keeping with the hypothesis that gain setting measurements for four anatomic regions are a source of bias. While the expert cohort demonstrates a clear trend towards consistency in gain adjustment, the novice cohort does not demonstrate any such consistency. Based on these findings we do not feel that any modification of our scoring algorithm is warranted to account for repeated gain sampling bias.

The second data point worthy of special consideration is the inclusion of the “Autonomy” domain of the GRS data set. This domain is included in the proposed scoring system as it has obvious potential benefits in assessing a cohort of new trainees. The nature of our current study however does not warrant an assessment of autonomy as no participants, novice or expert, were given any coaching or guidance during completion of their FAST assessments except the ability to watch the instructional video at the start of the trial. Assessment of the Autonomy domain can be performed in future prospective studies using the same GRS model.

Comparing hand motion analysis in our study versus others is more difficult due to the specific relationship between hand motion measurements and the procedure assessed; the number of movements or path length in our assessment of FAST imaging, for example, would not be comparable to the movements or path length in the aforementioned axillary nerve block studies. Because no other paper to date has studied hand motion measures in relation to FAST

imaging, it is not possible to compare our results directly against the literature findings. A summary of hand motion analysis in various technical skills is demonstrated in Table 28 below.

Table 28: Summary of Hand Motion Domain Validity Statistical Significance in the Reviewed Literature

Study	Procedure	Left Hand Path Length	Right Hand Path Length	Total Path Length	Left Hand Movements	Right Hand Movements	Total Movements	Time
Ours/Current*	FAST Ultrasound	Yes†	Yes	Yes	Yes	Yes	Yes	No
Chin et al., imaging	US Guided Nerve Block	/	Yes	/	/	Yes	/	Yes
Chin et al., “needling”	US Guided Nerve Block	Yes	Yes	/	Yes	Yes	/	Yes
Alvand et al.	Arthroscopic Meniscal Repair	/	/	Yes	/	/	Yes	Yes
Howells et al. 2008	Simulated Arthroscopy	/	/	Yes	/	/	Yes	Yes
Howells et al. 2009	Arthroscopic Suturing	/	/	Yes	/	/	Yes	Yes
Aggarwal et al.	Laparoscopic Cholecystectom y	/	/	No	/	/	No	Yes
Aggarwal et al.	Laparoscopic “Clip and Cut Duct”	/	/	No	/	/	No	Yes
Aggarwal et al.	Laparoscopic “Clip and Cut	/	/	No	/	/	No	Yes

	Artery”							
Aggarwal et al.	Laparoscopic Cystic Duct Dissection	/	/	Yes	/	/	Yes	Yes
Moorthy et al.	Simulated Laparoscopic Suturing	/	/	Yes	/	/	/	Yes
Study	Procedure	Left Hand Path Length	Right Hand Path Length	Total Path Length	Left Hand Movements	Right Hand Movements	Total Movements	Time
Xeroulis et al.	Laparoscopic Peg Transfer	/	/	Yes	/	/	Yes	Yes
Xeroulis et al.	Laparoscopic Pattern Cutting	/	/	No	/	/	No	Yes
Xeroulis et al.	Laparoscopic Endolooping	/	/	No	/	/	Yes	Yes
Xeroulis et al.	Laparoscopic Suturing	/	/	Yes	/	/	Yes	Yes
Hayter et al.	Epidural Catheter Insertion	/	/	Yes	/	/	Yes	Yes
Gallagher et al. 2002	Various Laparoscopic Object Manipulations	/	/	/	Yes	Yes	/	Yes

49,56,76,92,95-97,100,109

* Left, Right, and Total movements defined by “Velocity” modality

† Yes = statistically significant result reported, No = statistically insignificant result reported

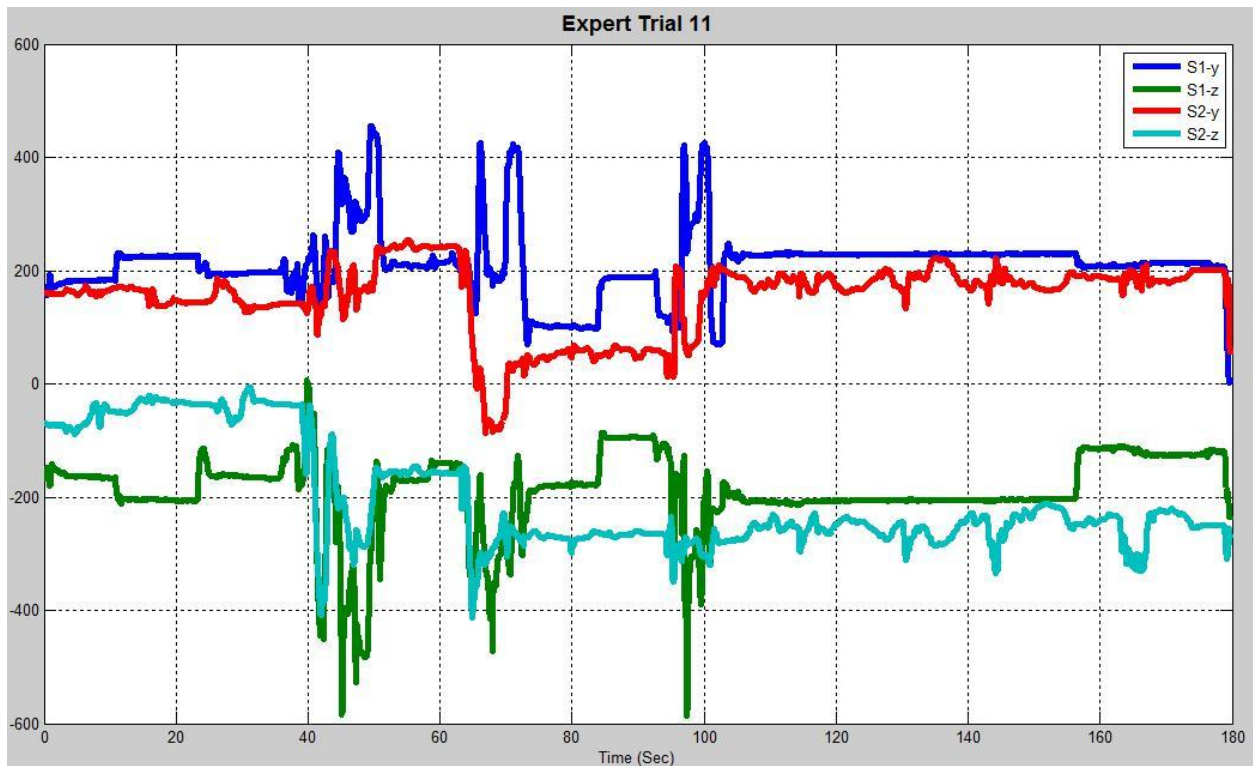
In summary, our own study demonstrates only one significant outlier from the reported literature. Every study which separated left and right hand movements for either path length (ours, Chin et al.) or number of movements (ours, Chin et al., and Gallagher et al.) found that each hand is an independent predictor of expertise. We can see that ten of fifteen studies conclude that path length is a significant predictor and eleven of fifteen studies conclude that number of movements is a significant predictor, both figures in agreement with our own findings.

The major outlier in our results is the lack of significance of the time variable. Our study carried with it one limitation in this regard. Our male volunteer who served as the ultrasound model for the twenty four trials in our study was a particularly thin male with a narrow costosternal angle which was a significant barrier to achieving a satisfactory subxiphoid view on the FAST pericardial exam. While our volunteer was not chosen expressly for this reason, his natural anatomic constraints provided an easy way to measure one item on the FAST Image Acquisition Checklist under the pericardium scoring which reads “Optimizes view of pericardium using adjuncts as necessary.” FAST trainees are introduced to multiple adjuncts they may use to optimize the pericardial view including breath holding, hip flexion, and the parasternal view as an alternate to the subxiphoid view. The use of the parasternal view was in fact included in our instructional video and our volunteer’s body habitus allowed us to adequately assess the participants’ use of adjuncts to achieve this view.

Unfortunately while this unexpected finding may have improved the fidelity of our pericardial view task checklist scores, it probably compromised our assessment of time. Many participants, particularly the expert cohort who were better versed in various adjuncts, spent considerable amounts of time assessing the pericardium via the un-ideal subxiphoid view

including the use of adjuncts in that view before finally attempting the parasternal view. This phenomenon is demonstrated in Figure 11, a graphical representation of hand motions as one expert participant performed the FAST exam.

Figure 11: Hand Motion Mapping of a Single Expert Participant's FAST Performance

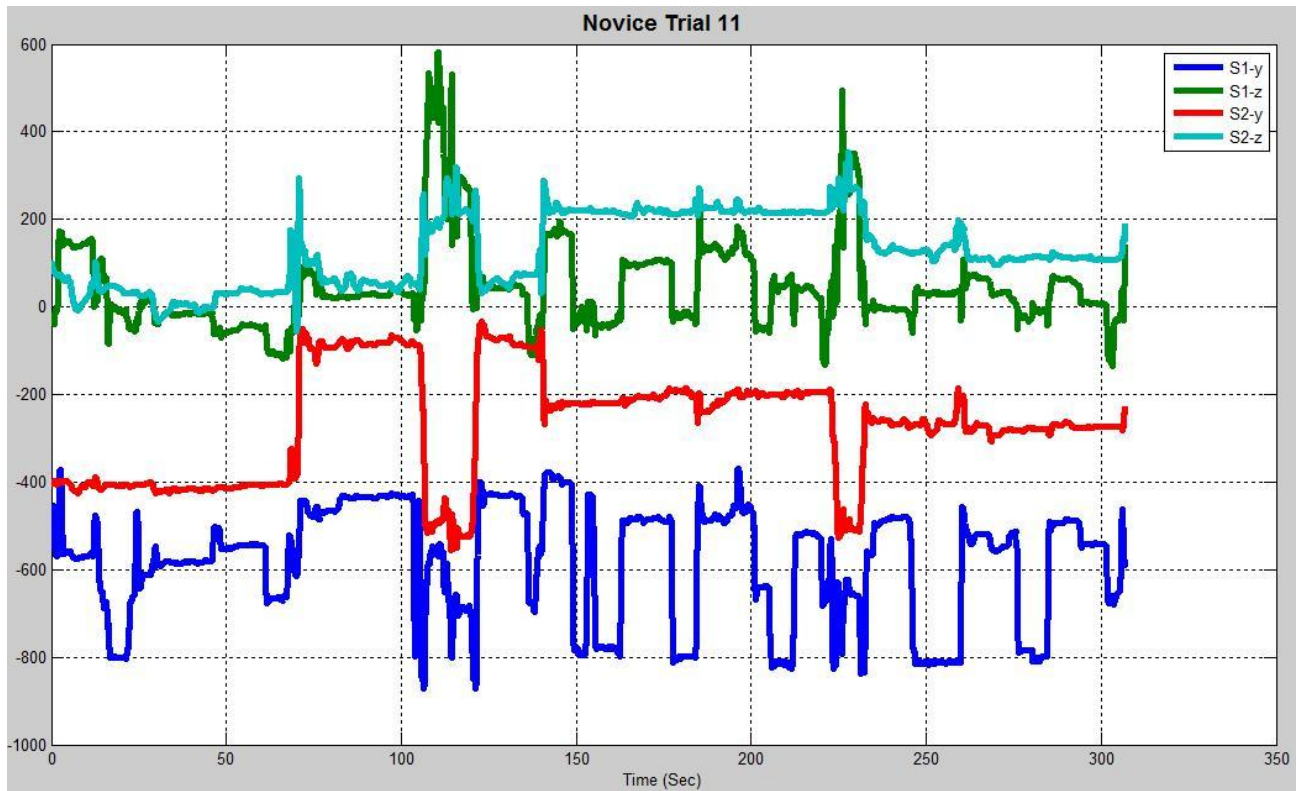


This graphical representation of hand movements clearly demonstrates four distinct phases of the FAST exam. From time 0 to approximately time 40 the participant achieves one view; from time 45 to time 63, approximately, he or she achieves the second and from time 66 to time 96 he or she achieves the third. The small time intervals between views where there are large amplitude movements represent the repositioning of the examiner's hands between anatomic regions to be imaged. The first three regions required this participant approximately 40, 18, and 30 seconds respectively. Notably however, approximately 77 seconds – from time 103 to the end of the study at 180 – are spent on the final region to be imaged. A single large,

distinct movement at approximately 155 seconds represents in this participant the choice to abandon a subxiphoid view and attempt a parasternal view, which quickly results in a satisfactory image.

A typical novice hand motion graph demonstrates significant differences between the novice and expert performance (Figure 12).

Figure 12: Hand Motion Mapping of a Single Novice Participant's FAST Performance



Unlike the expert performance demonstrated in Figure 11, the novice hand motion curves demonstrate considerably more high-amplitude movements at all times rather than simply during the transitions between anatomic regions indicating that the novice performers were incapable of the fine, smooth transitions expected of a competent sonographer. Inspection of the major transitions (consider the two green or one red line in Figure 12 above) shows that this performer

imaged five regions rather than four, with regions imaged from time 0 to 60, 65 to 105, 120 to 130, 140 to 220, and 230 to 300 seconds, suggesting unfamiliarity with the requisite views and where to position the probe to achieve them. Notably the final region imaged (the pericardial space) lacks a distinct repositioning which would suggest an adjunct view and, with the exception of region two, each region imaged was of approximately length of time in contrast to the expert performer.

Most novices did not attempt to use adjuncts despite their demonstration in our training video. While novices performed all four views slowly, experts performed the three abdominal views quickly and spent excessive time obtaining high quality, albeit slow, images of the pericardium. Novices on the other hand, spent a short time attempting the subxiphoid view and then often gave up and terminated the exam early without obtaining an adequate pericardial image. Had this trial included a more “anatomically average” volunteer, or multiple different volunteers, we suspect that the time variable of hand motion which *trended* towards significance would truly *reach* significance. Repeated trials in the future will be able to demonstrate the veracity of this assertion.

In the above summary of hand motion analysis data, our own data for number of movements was intentionally selected based on the “velocity” trigger. An individual movement is defined through processing software as the triggering of a predefined gateway. A movement may be counted for example when velocity changes by a predetermined value, when position changes by a predetermined value, when acceleration changes by a predetermined value, or when jerk changes by a predetermined value. We found that a velocity trigger was the only modality to reach statistical significance. In our comparison chart however this value was selected not because it is the only significant result, but because it is the modality widely supported by the

literature. The most widely used and researched hand motion device is the Imperial College Surgical Assessment Device (ICSAD), which is used by the majority of the previously cited studies. The ICSAD device itself calculates each movement by a velocity trigger. Dr. Dosis, one of the creators of the device, wrote in a 2005 Archives of Surgery paper describing the device that “An a priori condition stated that each movement is counted when the physical movement high-peak velocity value exceeds a predefined velocity tolerance.”¹⁵³ This modality of measuring movements was later validated by comparing ICSAD movement data to high fidelity video recordings and manually counting the motions witnessed¹⁵⁴. Our use of the velocity-gated movement calculation was thus based on the accepted and most widely documented standard in the literature.

Other modalities of movement measurement may in the future be validated as means of assessing participant movement but in our study the modalities of position, acceleration and jerk all failed to reach statistical significance. It is apparent from a cursory examination of our data that the sensitivity of the device to measure movement under these alternate definitions is not well calibrated. For example, the same device and same software measured total movement averages in the novice cohort of 73 (position), 452 (velocity), 601 (acceleration), and 709 (jerk). It is entirely possible that more work in the future to define these tolerances differently may lead to finding one or more other movement-definitions as significant, but it is clear from our data that our technology as calibrated in this study is not up to the task. As such, in future trials we will consider only the movements as measured by a velocity algorithm to be aligned with the existing literature and our own findings.

Finally, within our data of hand motion, we also measured a number of previously undescribed and non-validated measures using our device and software. There are no published

comparators for these items which included maximum and average values for velocity, acceleration and jerk, and also movements per second. Given the previous discussion of the unproven and questionable utility of the acceleration and jerk modalities, there is no reasonable conclusion that can be drawn from our data despite the finding that average jerk was apparently a significant predictor of expertise. The finding however of the significance of movements per second defined by velocity changes *is* in keeping with proven definitions in the literature and suggests that experts, who perform fewer average movements per second, may be demonstrating a more deliberate and careful imaging technique compared to novices who may be simply randomly “scanning” the model for any recognizable structures. Ultimately this is simply a re-expression of previously described data showing that experts perform significantly fewer total movements without a significant decrease in the time needed to complete the exam.

One potential source of error which may account for the failure to reach significance within these domains is the lack of calibration for each domain. Our hand motion apparatus was calibrated for change in position and time. These measures alone allow us to calculate velocity directly, but measures of acceleration and jerk require mathematical differentiation. Any underlying error in the initial calibration would thus be magnified by orders of magnitude for acceleration and jerk respectively, leading to substantial error. It is not surprising, therefore, that these measures are inconsistent in their statistical outcomes.

In summary, our findings are generally in keeping within the reported values in the comparable literature with the notable but explainable exception of the “time” variable.

Discussion: Implications of Outcomes and Future Proposals

We previously outlined the many proposals for the objective assessment of technical skills, including the use of global rating scales, task specific checklists and three dimensional hand motion tracking as we utilized here in addition to error analysis and simulation metrics. The multiple modalities of assessment all feature unique positive attributes which make them good assessment tools as well as weaknesses which need to be considered in study design. By combining multiple proven modalities we hope to overcome these individual shortcomings and create a strong, reliable, and valid model of assessment in ultrasound image acquisition skills.

Our study is based on multiple independent assessments: two task specific checklists, one global rating scale, and one hand motion assessment. These items could be considered independently or combined into a single assessment modality. We must consider both optimum mathematic models and also practical considerations in selecting the “best” possible predictive model. As such it seems fitting to divide the data into two proposed models:

The first model which warrants consideration is the overall best fitting model. This would be a multivariate model derived from all significant univariate results in our assessment: the FAST Image Acquisition total score, the FAST Representative Image total score, the FAST Global Rating Scale total score, the total path length traveled and the total number of movements defined by velocity. This model would ideally demonstrate the absolute maximum discriminatory power of the tools we have described and would serve as the baseline for comparison of our other more practical proposals.

While we have attempted to develop such a model to use as a baseline for comparison against all other less complete models, the collinearity of the variables was too high to generate

high quality multivariate models. While unanticipated, this results is, in hindsight, not surprising. All of our scoring systems included systematic overlapping of assessments. Both task checklist and GRS scores including elements of gain and depth. The GRS score also overlapped with the time aspect of hand motion analysis, and hand motion measures of number of movements and path length would certainly also be captured by the GRS assessments of flow, probe handling, image sweeping, and possibly initial probe placement. The resulting collinearity limits our ability to develop an “all encompassing” baseline model.

One intrinsic flaw of using such a comprehensive model, were it possible to develop, is that it requires the end user to purchase multiple pieces of hardware and possibly commission custom software to make sense of the hand motion data. A well validated and powerful scoring system is not useful if it is unwieldy or prohibitive to use. As such, a second model could be considered for practical purposes which excludes the hand motion analysis data. Removing this component from consideration would save prospective users of the scoring system thousands of dollars in equipment costs and multiple hours of software training; future downward cost adjustments may negate this as a barrier, but that remains to be seen.

In looking at our scoring systems’ performance it is clear that the question of collinearity will again be encountered. We minimized this by using only two of our three scoring systems. Because the GRS and task checklists measure different domains of performance, and inclusion of the GRS is a proven element in the OSATS style assessments, it is clear that a proposed two-score model must include the GRS score and one of the two task-checklist scores. Because of our goal of simplifying the scoring process and minimizing the need to purchase costly or prohibitive equipment, we elected to develop a model using the dynamic-image assessment of the Image Acquisition Checklist to remove the need for recording or saving specialized images;

in theory, this two-score checklist could be scored “live” without need for image review by the examiner post-FAST. The Image Acquisition Checklist is also more comprehensive than the Representative Image Checklist by design (the Representative Image Checklist being a simplified version of the Image Acquisition Checklist), which includes elements of the FAST exam not captured in static imaging. We therefore attempted multivariate modeling using the GRS score and the Image Acquisition Score; unfortunately, though the problem was dramatically reduced from what was encountered on the “all encompassing” model design, we again were limited by collinearity of the variables.

Overcoming collinearity by scoring system design may have been a possibility in retrospect, but would not likely have changed the actual structure of the scoring systems fundamentally. By design, the GRS and TSC systems are assessing distinctly different elements of the same phenomena; the GRS is a quality assessment of general skills while the TSC is a binary assessment of specific tasks. Collinearity results from the high correlation expected between a high quality performance and the successful achievement of goals. Phrased another way, participants who are performing a high quality exam (and thus scoring highly on the GRS) are, intuitively, also the ones more likely to capture all of the critical views (scoring highly on the TSC). While we could attempt to collapse the GRS and TSC scores into a unified summative score, the design of the scoring system would still benefit from both a multi-level quality assessment component and a binary imaging success component; functionally therefore, a condensed unified scoring system would be nearly identical to our two-score system proposed here.

This, however, does not necessarily limit our ability to develop a predictive model using these two scores. As demonstrated from our results and our ROC curve modeling, the predictive

capabilities of these scores based on area under the ROC curve is 89.9% and 97.6% respectively for the Image Acquisition and GRS total scores. The best “passing” score for each scoring system is also demonstrated in our results, with a score of 15 on the Image Acquisition Checklist corresponding to a 60.0% probability of expert status and a score of 25 on the GRS corresponding to a 56.9% probability of expert status. These numbers were derived from within-sample analysis as we did not record an out-of-sample dataset; we feel, however, that the calculated sensitivity and specificity represent close approximations of the true values. In scoring future participants, a score of equal to or greater than 15 and 25 on the respective studies would be scored as a “pass” based on this data, while a single score below the defined threshold (even when achieving a passing score on the other test) would result in a “fail” assessment.

Using a streamlined model composed of the GRS and Image Acquisition scores we can objectively assess ultrasound imaging capabilities and make predictions about an individual’s probability of being expert certified. Currently, the criteria for certification of ultrasound competency vary by the certifying body and geographic locations and demonstrate no evidence-based reasoning for the divergent certification criteria required. Armed with a validated, reliable, and objective measure of ultrasound imaging skills we have the ability to develop evidence-based guidelines for awarding certification of competency in FAST ultrasound.

Establishing certification using our construct could be achieved by using our two-score model to evaluate the learning curve of the procedure. A prospective cohort of trainees could be serially examined with their success (pass versus failure) rate plotted as a learning curve. At some point, as yet undefined by the literature, their scores and capabilities would be expected to reach a plateau beyond which further practise would yield minimal improvement. It would be reasonable then to define the minimum training standard based on such data, wherein trainees are

expected to have enough experience that they are beyond the expected inflection point of the learning curve and thus have maximized or nearly maximized their skills before being certified as a solo practitioner. The caveat with this means of assessment is two-fold. First, if the learning curve were to suggest a large number – one hundred or more studies, for example – then both the study mapping the learning curve, as well as the rigorous training requirements suggested by the model are impractical. This is unlikely given the current standard which allows certification in this skill set after only fifty studies in many regions. Second, it is possible to imagine that an individual may reach their maximum potential (ie. a plateau in their learning curve) but still fail to demonstrate competence. To guard against this, the learning curve model should be paired with some objective measurement of skill such as a minimum score which must be achieved by the end of the learning period.

The most common method of designing such a learning curve relies on a predefined pass or fail threshold which converts scores into a binary model; trainees score a one for a passing performance and a zero for a failing performance. Having established a binary outcome, a prospective cohort of trainees could be enrolled to perform serial assessments and subjected to a cumulative sum (CUSUM) analysis. A CUSUM analysis allows us to measure the learning curve of an individual within the constraints of a specified acceptable failure rate. Allowing a higher failure rate for example would allow us to determine “competency” with a lesser degree of experience, whereas demanding a lower failure rate will require a higher number of repetitions in order to perfect skills¹⁵⁵. After choosing both passing criteria and an acceptable failure rate the CUSUM technique could be used to map out a learning curve of participants and plot the rate at which FAST ultrasound imaging skill is acquired, to give us an impression of how many trials or how much experience should be required before certifying any trainee as competent. Passing

or failing an individual trial can be assessed based on the results of our current study, with passing performance indicated by an Image Acquisition Checklist score greater than or equal to 15 and a Global Rating Scale total score greater than or equal to 25.

We have suggested the use of the ROC inflection points as a potential definition of a passing benchmark, but other means of identifying a passing criteria have been proposed.

Some authors arbitrarily choose a minimum score required to pass; this must be accompanied by a conscious anchoring of the global rating scale components so that a known value on the Likert scale represents a passing performance and higher scores represent exceptional performance¹⁵⁶. Using this criteria in our model, we would demand a score of “4” in each GRS domain before scoring that performance as a “pass;” no comparable value is defined for the task checklists or hand motion data. The selection of a “reasonable” but otherwise arbitrary passing score has been used, for example, by Nair et al. in their assessment of cardiology fellows’ echocardiography performance³⁸. Here, an arbitrary score of 60% of the maximum possible score was determined to be a “reasonable” passing grade and the same practice could be applied to our measures; this, however, would fail to satisfy a reasonable definition of “competent” and would fall within the classification of “arbitrary.” Another proposal would see participants in a tiered program be required to demonstrate their skills as equal or better than 75% of the scores achieved by the tier above them; that is, to advance from a “level two” to a “level three” trainee, a level two trainee must score 75% of what a level three would on the same task¹⁵⁷. This model is not easily adapted to FAST performance as we do not recognize various tiers in FAST skill sets; there is no super-certified cohort to compare to at present, though that may be an avenue to explore as the test evolves in the future.

It may be possible to define the optimum passing score with further research. One example of a super-certified cohort for comparison might be comparison against image acquisition skills of radiologists or certified ultrasound technologists. A non-expert seeking FAST certification might then be required to demonstrate a lesser degree of proficiency while meeting minimum requirements, similar to the tiered-training approach described previously. For example, a passing FAST performance may be defined as scoring 75% of what a professional radiologist achieves in his or her assessment of the same patient on the same day, though the passing grade remains arbitrarily chosen. Another approach might be to perform the scoring assessment on existing, known, certified experts as we have in our study and compare their scores to those of newly minted experts who have just completed their first training course. With training fresh in their memory and dozens of practise sessions over the course of a short time interval during their certification, we might hypothesize that the newly-certified individuals would score higher than the existing experts; in that case, we might suggest that a “passing” grade for a new trainee must not only meet but *exceed* the scores demonstrated by the expert cohort in our study.

Consideration of the other proposals for determining a passing grade discussed in the literature ultimately demonstrates that most such concepts are arbitrary, subjective, and without sufficient evidence to support them. Our own data provides not only a proposed passing grade, but the ability to derive predictions from it based on actual observations and evidence. Moving forward, in spite of the other proposals discussed or used in the ultrasound literature, we believe that the best determination of a passing performance is the evidence-guided ROC-curve derived scores demonstrated in this study.

Traditional interpretation of an ROC curve requires some subjective modification of the data. Procedures or tests with highly important outcomes may have their “pass” marks modified to minimize potential risks. For example, the failure tolerance in interpreting an electrocardiogram in a patient with chest pain must be lower than the failure tolerance of assessing a joint for osteoarthritis; failure in one may lead to a fatal missed diagnosis while failure in the other contributes largely to discomfort. We must consider that the potential harm of a falsely negative FAST exam may have severe negative consequences for the critically ill trauma patient and strive to set passing performance at such a level as to minimize those risks. As such, one could argue that the “pass” grade should be adjusted to a *higher* score to *increase* the probability that individuals obtaining such a score are indeed experts before being certified as such.

This modification seems unnecessary in our proposal for two reasons. First, no study has yet demonstrated a correlation between quality of FAST imaging performance and clinical outcomes. While we might intuit a relationship between poor imaging technique and patient harm, we are unable to quantify or prove this concept and so pass-threshold modification based on this premise would be entirely subjective and potentially biased. Second, pass-threshold modification based on our ROC curves would be the wrong way to reduce harm. It is unreasonable to award expert status based on the performance of a single FAST exam, even when performance is exemplary. Examination of multiple exams must still be the standard for certification, the only question being “how many?” Rather than adjusting the pass threshold of our GRS or Image Acquisition scores, the ideal means of risk-reduction is to define a learning curve with a lower acceptable fail rate. In this way we are not demanding exemplary

performance on every individual exam to award certification, but instead are demanding *consistently good* performance from candidates to earn credentials.

What is left, therefore is to score a separate cohort of FAST practitioners to validate the findings of this study including the predictive capacity of our two-score model, and to serially examine a prospective cohort of trainees to define the FAST learning curve.

Serial performances by trainees can be scored as a pass or fail based on our ROC curves for the GRS and Image Acquisition scores and subjected to a CUSUM analysis to determine when performance becomes consistent enough that it meets our requirements. By using our model to determine the probability of expertise and scoring trials accordingly as a pass or failure, we can map out the learning curve for FAST imaging using the CUSUM technique and have both a *theoretical* mean number of trials for all individuals to become an expert (based on the learning curves) and an *individual* means of measuring a specific trainee's performance. Using this information we could provide an evidence-based guideline for FAST ultrasound certification: participants must complete a minimum number of studies as determined by the CUSUM analysis and meet personal scores consistent with expertise based on our model.

Overall our study design could be improved in several ways. Firstly, a larger cohort of both novices and experts would have yielded a greater fidelity of statistical analysis. While this consideration applies to nearly every study conducted, there are obviously benefits in our particular study to increasing enrolment or repeating the study in the future with a larger cohort. While the majority of our broad categories demonstrated statistically significant results with the notable exceptions of the pelvic space on the Representative Image Checklist ($p=0.1454$) and Image Acquisition Checklist ($p=0.0500$), and the time assessment on hand motion analysis ($p=0.1219$), we lacked statistical power to determine which individual checklist items

corresponded strongly with expert status. It is entirely possible that specific items within those broad categories are themselves strong predictors of expertise and could conceivably allow us to streamline the assessment process by providing more weight to those items. Without a larger cohort it is difficult to make these determinations.

A larger cohort of expert *reviewers* would have allowed us to narrow the confidence intervals on our inter-rater reliability assessments. While the current data with only two reviewers demonstrates substantial agreement between reviewers, greater weight and confidence would be provided by having more experts involved in the review of and scoring of our recorded trials. Furthermore, defining anchors at all five points of the Likert scale in each domain may have improved inter-rater reliability and improved the kappa results, by clearly defining what criteria represent the previously-undefined scores of “2” and “4.” One final manoeuvre to potentially improve the agreement would be to compare the agreement of reviewers who are recognized as super experts. Our reviewers in this study were each experts with identical training to the cohort being assessed; the observed heterogeneity of performance could therefore imply heterogeneity of scoring. If the expert reviewers were individuals with beyond-standard FAST or ultrasound training, we may find improved agreement in their interpretation of the images generated.

Our study was conducted using a single male volunteer who was subjected to all forty-eight ultrasound investigations. This poses three potential problems: first, we have not assessed any of our cohort on their interpretation of or imaging of female anatomy, which is most different in our weakest area of assessment, the pelvis. With the added anatomic considerations of the pelvis (specifically the visible uterus behind the bladder) we might expect scores on this quadrant to differ from those in our current study. We hypothesize that the use of a female

volunteer may favor better performance by the expert cohort who would likely have the clinical experience to recognize the need to change the depth of imaging to ensure assessment of the space *behind* the bladder, while novices may not intuit that fact immediately. Second, our volunteer was a healthy male with an absence of pathology. While not a limitation per se, as our study's aim was to evaluate imaging technique rather than clinical findings, potential criticism of the study may arise from a failure to demonstrate that performance correlates with increased identification of pathology. Finally, as previously described, our male volunteer had specific anatomic constraints which made imaging of the subxiphoid pericardial space quite challenging if not impossible, requiring the use of adjuncts including alternate views. The use of adjuncts and alternate views are of course a critical component of mastering FAST imaging, but the uniqueness of the volunteer may have been responsible for the statistically insignificant findings with regards to the time assessment on hand motion analysis.

A final consideration would be to compare the scores of the community of experts to the scores of newly certified experts. Many credentialed training enterprises in medicine require ongoing practise and/or periodic recertification. Courses such as the Basic Life Support (BLS), Advanced Cardiac Life Support (ACLS), and the Advanced Trauma Life Support (ATLS) courses all require recertification after several years. FAST imaging, at this point, does not. One might hypothesize that the use of experts who have *not* recently been certified might contribute to overall lower scores and worse performance by the expert cohort. If our study had been negative and our scoring systems had been unable to discern a difference between the expert and novice cohorts, this would be a larger concern. As such however, the greater question is this: *if* experts' skills do regress and worsen as time passes from their initial training, and we base our expert-certification criteria on the study of remotely certified experts, are we in fact *lowering* the

threshold of skills in our training cohort? That is, if we offer credentials to trainees based on meeting the skills of experts who have *decreased* in measurable skill level, will that cohort *also* decrease in measured skills and ultimately be less skilled than the experts we had originally asked them to meet or best? Clearly this is an area which will require further research before credentialing proposals based on our work can be put forward.

We are also left with highly discriminatory logistic regression models for two of the three previously validated measures of hand motion analysis. While these items are not integrated into a proposed scoring system, for the purposes of simplifying the proposed credentialing metric, there is still potential use and value in these findings. The best way to apply the validated hand motion findings of this study is to integrate such metrics into a high fidelity simulator. Simulators which exist for surgical techniques and ultrasound already measure metrics of hand motion in many cases. Assessment of performance using our validated metrics on a simulation device would provide a nearly instant assessment of a trainee's performance on a variety of exams which, when combined with the intrinsic ability of a simulator to determine if critical views have been obtained, would provide to trainees objective assessments of both success at imaging and efficiency of technique. One distinct advantage of this approach is that training tools via simulation can be assessed as a standard format across multiple regions allowing comparisons between groups trained in different regions on the exact same model with identical pathologies; this may be a means of comparing training techniques to compare outcomes which might otherwise be limited by the heterogeneity of anatomic models.

The future work to develop a credentialing process using the probability of expertise and CUSUM learning curve derivation is only one potential avenue of research which may stem from this project. As described, our algorithm could be used to measure a difference in performance

between newly-credentialed performers and their in-practise colleagues to document whether or not there is a regression of skills. This might ultimately lead to an evidence-based argument for periodic recertification to maintain FAST skills and credentials. We might also use this scoring system to compare the learning curves of surgical trainees versus emergency medicine trainees as a study less on FAST itself but on procedures-based training for medical trainees, hypothesizing that the surgical cohort (being a predominantly procedural specialty) would master the skill with fewer repetitions than the ER medicine cohort. This work may also serve as a tool to improve FAST training in the future. Does performing fifty ultrasounds back-to-back in the context of a training course provide the same level of experience as the performance of fifty clinically-indicated studies in the emergency room, for example? We now have a validated tool to help answer that question by measuring the skill sets of cohorts trained in different ways. Finally, as previously suggested, we might attempt to answer the question of whether or not quality of FAST imaging correlates with clinically significant outcomes such as the rate of false negative, false positive, or simply non-diagnostic tests. In theory, if quality of FAST imaging performance were found *not* to correlate with outcomes, it might inspire the question of “what is the value of certification?”

Comparing the quality of a FAST exam with clinical outcomes might be achieved in one of several ways. One method would be to assess (with our scoring systems) a local cohort of FAST practitioners as a baseline and to note their scores. Comparison of FAST outcomes to gold-standard diagnostic techniques (such as CT scan or findings at laparotomy) should theoretically demonstrate higher sensitivity and specificity of the FAST exams performed by those practitioners with higher baseline scores. It may be possible to simplify this proposed study by way of simulation; a patient’s abdomen might be serially infused with an increasing,

known quantity of fluid (using for example a peritoneal dialysis patient). Practitioners with known FAST quality assessment scores could be asked to repeatedly assess the patient with changing volumes of intraperitoneal fluid to determine whether sensitivity and specificity of their performance for a clinically relevant outcome (detection of known free fluid) is correlated to baseline scores. Both of these proposals would compare performance scores to radiologic findings; it could be argued however that a clinically relevant amount of intraperitoneal fluid would be detected even with a poor performance and that the higher theoretical sensitivity and specificity would not impact patient outcomes. One purely clinical approach would be to compare whether the baseline quality scores of a cohort of practitioners correlates to reduced times from presentation to intervention in an operating theatre or angiography suite. Such a study would have to be quite large to account for the variety of injuries and heterogeneity of hemodynamic stability within the patient cohort, but could theoretically demonstrate that improved imaging skill makes for tangible improvements in patient care and outcomes.

Finally and (perhaps) most importantly this study acts as a proof of concept that ultrasound imaging skills can be objectively assessed and quantified. It will be possible to revise FAST training criteria based on this work to follow a better evidence-based training protocol. Further, this study is easily replicated to other ultrasound-based fields of study. The evidence-based credentialing process may be extended to bedside echocardiography, extended FAST, pleural fluid aspiration, central line insertion, or any other modality using the ultrasound probe. As medicine becomes more segmentally divided into more and more discreet specialties, proof of competence is likely to become a larger and larger issue for physicians. This study puts us at the forefront of developing an evidence-based system of documenting expertise in ultrasound-based technical skill.

Conclusion

The proposed scoring system using the Global Rating Scale and Image Acquisition Checklist meets multiple definitions of validity and reliability, in addition to being feasible as demonstrated by the lack of requisite technical recording or assessment tools. The overlapping assessments between our global rating scale, checklists, and hand motion assessments provide a reassuring measure of convergent and divergent validity which contributes to overall construct validity. Performance of the tasks on a live human volunteer with a real ultrasound device assures us of content validity, and comparison of formally trained and certified experts against untrained novices provides a measure of construct validity. Our inter-rater assessments demonstrate the moderate to substantial reliability of the assessment tools, achieving our goal of developing a valid, reliable, and feasible assessment tool for the objective assessment of FAST ultrasound skill.

This project represents the first step in what may ultimately redefine FAST ultrasound training standards, with the potential to influence the training standards for many other ultrasound and non-ultrasound point of care procedures.

References

1. Tse F, Barkun JS, Romagnuolo J, Friedman G, Bornstein JD, Barkun AN. Nonoperative imaging techniques in suspected biliary tract obstruction. *HPB : the official journal of the International Hepato Pancreato Biliary Association*. 2006 Jan;8(6):409–25.
2. Wasnik AP, Menias CO, Platt JF, Lalchandani UR, Bedi DG, Elsayes KM. Multimodality imaging of ovarian cystic lesions: Review with an imaging based algorithmic approach. *World Journal of Radiology*. 2013 Mar 28;5(3):113–25.
3. Fox JC, Bertoglio KC. Emergency Physician Performed Ultrasound for DVT Evaluation. *Thrombosis*. 2011 Jan;2011:938709.
4. Nelson BP, Chason K. Use of ultrasound by emergency medical services: a review. *International Journal of Emergency Medicine*. 2008 Dec;1(4):253–9.
5. Talley B, Adit A, Ali S, Ashley F, Janice A, Carlos A. Variable access to immediate bedside ultrasound in the emergency department. *Western Journal of Emergency Medicine*. 2011;12(April 2009).
6. Bolondi L. Message from the president. *European Journal of Ultrasound*. 1997;5(2):63–4.
7. Sidhu HS, Olubaniyi BO, Bhatnagar G. Role of Simulation-Based Education in Ultrasound Practice Training. *Journal of Ultrasound Medicine*. 2012;(31):785–91.
8. Blaivas M, Theodoro D, Sierzenski PR. Elevated intracranial pressure detected by bedside emergency ultrasonography of the optic nerve sheath. *Academic emergency medicine : Official Journal of the Society for Academic Emergency Medicine*. 2003 Apr;10(4):376–81.
9. Kirkpatrick AW, Sirois M, Laupland KB, Liu D, Rowan K, Ball CG, et al. Hand-Held Thoracic Sonography for Detecting Post-Traumatic Pneumothoraces: The Extended Focused Assessment With Sonography For Trauma (EFAST). *The Journal of Trauma: Injury, Infection, and Critical Care*. 2004 Aug;57(2):288–95.
10. Comert SS, Caglayan B, Akturk U, Fidan A, Kiral N, Parmaksız E, et al. The role of thoracic ultrasonography in the diagnosis of pulmonary embolism. *Annals of Thoracic Medicine*. 2013 Apr;8(2):99–104.
11. Zengin S, Al B, Genc S, Yildirim C, Ercan S, Dogan M, et al. Role of inferior vena cava and right ventricular diameter in assessment of volume status: a comparative study: Ultrasound and hypovolemia. *The American Journal of Emergency Medicine*. Elsevier Inc.; 2013 May;31(5):763–7.

12. Scalea TM, Rodriguez A, Chiu WC, Brenneman FD, Fallon WF, Kato K, et al. Focused Assessment with Sonography for Trauma (FAST): results from an international consensus conference. *The Journal of Trauma*. 1999 Mar;46(3):466–72.
13. Wherrett L, Boulanger B, McLellan B, Brenneman F, Rizoli S, Culhane J, et al. Hypotension After Blunt Abdominal Trauma: the role of emergent abdominal sonography in surgical triage. *Journal of Trauma*. 1996;41(5):815–20.
14. Thomas B, Falcone R, Vasquez D, Santanello S, Townsend M, Hockenberry S, et al. Ultrasound Evaluation of Blunt Abdominal Trauma: program implementation, initial experience, and learning curve. *Journal of Trauma*. 1997 Mar;42(3):384–90.
15. Ma OJ, Mateer JR, Ogata M, Kefer MP, Wittmann D, Aprahamian C. Prospective analysis of a rapid trauma ultrasound examination performed by emergency physicians. *The Journal of Trauma*. 1995. p. 879–85.
16. Rozycki GS, Ballard RB, Feliciano D V, Schmidt J a, Pennington SD. Surgeon-performed ultrasound for the assessment of truncal injuries: lessons learned from 1540 patients. *Annals of Surgery*. 1998 Oct;228(4):557–67.
17. Goodman J. Some fundamental properties of speckle. *Journal of the Optical Society of America*. 1976;66(11):1145–50.
18. Ollerton JE, Sugrue M, Balogh Z, D'Amours SK, Giles a, Wyllie P. Prospective study to evaluate the influence of FAST on trauma patient management. *The Journal of Trauma*. 2006 Apr;60(4):785–91.
19. Udobi KF, Rodriguez a, Chiu WC, Scalea TM. Role of ultrasonography in penetrating abdominal trauma: a prospective clinical study. *The Journal of Trauma*. 2001 Mar;50(3):475–9.
20. Boulanger B, McLellan B, Brenneman F, Ochoa J, Kirkpatrick A. Prospective evidence of the superiority of a sonography-based algorithm in the assessment of blunt abdominal trauma. *Journal of Trauma*. 1999;47(4):632.
21. Forster R, Pillasch J, Zielke A, Malewski U, Rothmund M. Ultrasonography in blunt abdominal trauma: Influence of the investigator's experience. *Journal of Trauma*. 1993;34(2):264–9.
22. Dawoud D, Lyndon W, Mrug S, Bissler JJ, Mrug M. Impact of ultrasound-guided kidney biopsy simulation on trainee confidence and biopsy outcomes. *American Journal of Nephrology*. 2012 Jan;36(6):570–4.
23. Padman M, Warwick AM, Fernandes JA, Flowers MJ, Davies AG, Bell MJ. Closed reduction and stabilization of supracondylar fractures of the humerus in children: the

- crucial factor of surgical experience. *Journal of Pediatric Orthopedics Part B*. 2010 Jul;19(4):298–303.
24. Gearhart SL, Wang M-H, Gilson MM, Chen B, Kern DE. Teaching and assessing technical proficiency in surgical subspecialty fellowships. *Journal of Surgical Education*. Elsevier Inc.; 2012;69(4):521–8.
 25. Nasca TJ, Day SH, Amis ES. The New Recommendations on Duty Hours from the ACGME Task Force. *New England Journal of Medicine*. 2010;3(1):1–6.
 26. PARIM Board of Directors. PARIM Collective Agreement 2011 to 2014. 2011 p. 1–48.
 27. MacRae HM. Objective assessment of technical skill. *The Surgeon : Journal of the Royal Colleges of Surgeons of Edinburgh and Ireland*. Elsevier Ltd; 2011 Jan;9 Suppl 1:S23–5.
 28. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. *American Journal of Surgery*. Elsevier Inc.; 2011 Oct;202(4):469–480.e6.
 29. Van Eaton EG, Tarpley JL, Solorzano CC, Cho CS, Weber SM, Termuhlen PM. Resident education in 2011: three key challenges on the road ahead. *Surgery*. Mosby, Inc.; 2011 Apr;149(4):465–73.
 30. Royal College of Physicians and Surgeons of Canada. Competency-based Medical Education. 2010.
 31. Swing SR. The ACGME outcome project: retrospective and prospective. *Medical Teacher*. 2007 Sep;29(7):648–54.
 32. Reznick R. Teaching and Testing Technical Skills. *American Journal of Surgery*. 1993;165(3):358–61.
 33. Darzi a, Mackay S. Assessment of surgical competence. *Quality in Health Care : QHC*. 2001 Dec;10 Suppl 2(Suppl II):ii64–9.
 34. Paget M, Wu C, McIlwrick J, Woloschuk W, Wright B, McLaughlin K. Rater variables associated with ITER ratings. *Advances in Health Sciences Education : theory and practice*. 2012 Jul 10.
 35. McLaughlin K, Bates J, Konkin J, Woloschuk W, Suddards C a, Regehr G. A comparison of performance evaluations of students on longitudinal integrated clerkships and rotation-based clerkships. *Academic Medicine : Journal of the Association of American Medical Colleges*. 2011 Oct;86(10 Suppl):S25–9.
 36. Hasnain M, Connell KJ, Downing SM, Olthoff A, Yudkowsky R. Toward meaningful evaluation of clinical competence: the role of direct observation in clerkship ratings.

- Academic Medicine : Journal of the Association of American Medical Colleges. 2004 Oct;79(10 Suppl):S21–4.
37. Reed WP, Kilkenny JW, Dias CE, Wexner SD. A prospective analysis of 3525 esophagogastroduodenoscopies performed by surgeons. *Surgical Endoscopy*. 2004 Jan;18(1):11–21.
 38. Nair P, Siu SC, Sloggett CE, Biclar L, Sidhu RS, Yu EHC. The assessment of technical and interpretative proficiency in echocardiography. *Journal of the American Society of Echocardiography* : official publication of the American Society of Echocardiography. 2006 Jul;19(7):924–31.
 39. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. *American Journal of Surgery*. 2007 May;193(5):551–5; discussion 555.
 40. Newell R, Burnard P. *Research for Evidence-Based Practice in Healthcare Second Edition*. Hoboken, NJ: Wiley-Blackwell; 2011. p. 150.
 41. Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. 2011. p. 17.4.1.
 42. Bonis P. Glossary of common biostatistical and epidemiological terms. UpToDate. 2012.
 43. Messick S. *Validity of Test Interpretation and Use*. Educational Testing Service. Princeton, NJ: Educational Testing Service; 1990. p. 33pp.
 44. Colman A. Convergent Validity. *A Dictionary of Psychology 3rd Edition*. 2012.
 45. Colman A. Discriminant Validity. *A Dictionary of Psychology, 3rd Edition*. 2012.
 46. Nevo B. Face Validity Revisited. *Journal of Educational Measurement*. 1985;22(4):287–93.
 47. Brothers TE, Wetherholt S. Importance of the faculty interview during the resident application process. *Journal of Surgical Education*. 2007;64(6):378–85.
 48. Ringsted C, Østergaard D, Ravn L, Pedersen JA, Berlac PA, Van der Vleuten CPM. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. *Medical Teacher*. 2003 Nov;25(6):654–8.
 49. Alvand A, Logishetty K, Middleton R, Khan T, Jackson WFM, Price AJ, et al. Validating a global rating scale to monitor individual resident learning curves during arthroscopic knee meniscal repair. *Arthroscopy* : the journal of arthroscopic & related surgery : Official Publication of the Arthroscopy Association of North America and the

- International Arthroscopy Association. Arthroscopy Association of North America; 2013 May;29(5):906–12.
50. Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD. The development of an objective model to assess arthroscopic performance. *The Journal of Bone and Joint Surgery American Volume*. 2009 Sep;91(9):2287–95.
 51. Koehler RJ, Amsdell S, Arendt E a, Bisson LJ, Bramen JP, Butler A, et al. The Arthroscopic Surgical Skill Evaluation Tool (ASSET). *The American Journal of Sports Medicine*. 2013 Jun;41(6):1229–37.
 52. Van Heest A, Putnam M, Agel J, Shanedling J, McPherson S, Schmitz C. Assessment of technical skills of orthopaedic surgery residents performing open carpal tunnel release surgery. *The Journal of Bone and Joint Surgery American Volume*. 2009 Dec;91(12):2811–7.
 53. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *Journal of the American College of Surgeons*. 2007 Feb;204(2):308–13.
 54. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *American Journal of Surgery*. 2005 Jul;190(1):107–13.
 55. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Annals of Surgery*. 2008 Feb;247(2):372–9.
 56. Moorthy K, Munz Y, Dosis a, Bello F, Chang a, Darzi a. Bimodal assessment of laparoscopic suturing skills: construct and concurrent validity. *Surgical Endoscopy*. 2004 Nov;18(11):1608–12.
 57. Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surgical Endoscopy*. 2004 Dec;18(12):1800–4.
 58. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *The British Journal of Surgery*. 1997 Feb;84(2):273–8.
 59. Hance J, Aggarwal R, Stanbridge R, Blauth C, Munz Y, Darzi A, et al. Objective assessment of technical skills in cardiac surgery. *European Journal of Cardio-Thoracic Surgery : Official Journal of the European Association for Cardio-thoracic Surgery*. 2005 Jul;28(1):157–62.

60. 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*. American Academy of Ophthalmology; 2009 Feb;116(2):257–62.
61. Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *American Journal of Obstetrics and Gynecology*. 2006 Aug;195(2):617–21; discussion 621–3.
62. Goff B, Lentz G, Lee D, Houmard B, Mandel L. Development of an Objective Structured Assessment of Technical Skill for Obstetrics and Gynecology Residents. *Obstetrics and Gynecology*. 2000;96(1):146–50.
63. Friedman Z, Katznelson R, Devito I, Siddiqui M, Chan V. Objective assessment of manual skills and proficiency in performing epidural anesthesia--video-assisted validation. *Regional Anesthesia and Pain Medicine*. 2006;31(4):304–10.
64. Khaliq T. Reliability of Results Produced Through Objective Structured Assessments of Technical Skills (OSATS) For Endotracheal Intubation. *Journal of the College of Physicians and Surgeons of Pakistan*. 2013;23(1):51–5.
65. Vassiliou MC, Kaneva PA, Poulouse BK, Dunkin BJ, Marks JM, Sadik R, et al. Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surgical Endoscopy*. 2010 Aug;24(8):1834–41.
66. Fanning J, Fenton B, Johnson C, Johnson J, Rehman S. Comparison of teenaged video gamers vs PGY-I residents in obstetrics and gynecology on a laparoscopic simulator. *Journal of Minimally Invasive Gynecology*. AAGL; 2011;18(2):169–72.
67. Regehr G, MacRae H. 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.
68. Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Academic Medicine : Journal of the Association of American Medical Colleges*. 2001 Oct;76(10):1053–5.
69. Hodges B, Regehr G. OSCE checklists do not capture increasing levels of expertise. *Academic Medicine*. 1999;74(10):1129–34.
70. Carbine DN, Finer NN, Knodel E, Rich W, Objective A. Video recording as a means of evaluating neonatal resuscitation performance. *Pediatrics*. 2013;106(4):654–8.
71. Van der Heide P a, Van Toledo-Eppinga L, Van der Heide M, Van der Lee JH. Assessment of neonatal resuscitation skills: a reliable and valid scoring system. *Resuscitation*. 2006 Nov;71(2):212–21.

72. Barrington MJ, Wong DM, Slater B, Ivanusic JJ, Ovens M. Ultrasound-guided regional anesthesia: how much practice do novices require before achieving competency in ultrasound needle visualization using a cadaver model. *Regional Anesthesia and Pain Medicine*. 2012;37(3):334–9.
73. De Oliveira Filho GR, Helayel PE, Da Conceição DB, Garzel IS, Pavei P, Ceccon MS. Learning curves and mathematical models for interventional ultrasound basic skills. *Anesthesia and Analgesia*. 2008 Feb;106(2):568–73.
74. Reznick R, Regehr G, MacRae H, McCulloch W. Testing Technical Skill via an Innovative “Bench Station” Exam. *American Journal of Surgery*. 1997;173(3):226–30.
75. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. *Academic Medicine*. 1996;71(12):1363–5.
76. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Pappas P, Dosis A, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Annals of Surgery*. 2007 Jun;245(6):992–9.
77. Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surgery Today*. 2013 Mar;43(3):271–5.
78. Francis HW, Masood H, Chaudhry KN, Laeeq K, Carey JP, Della Santina CC, et al. Objective assessment of mastoidectomy skills in the operating room. *Otology & neurotology* : official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology. 2010 Jul;31(5):759–65.
79. 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: Journal of the Society for Simulation in Healthcare*. 2009 Jan;4(1):6–16.
80. Vergis A, Gillman L, Minor S, Taylor M, Park J. Structured assessment format for evaluating operative reports in general surgery. *American Journal of Surgery*. 2008 Jan;195(1):24–9.
81. Gillman LM, Vergis A, Park J, Minor S, Taylor M. Structured operative reporting: a randomized trial using dictation templates to improve operative reporting. *American Journal of Surgery*. Elsevier Inc.; 2010 Jun;199(6):846–50.

82. Bann S, Khan M, Datta V, Darzi A. Surgical skill is predicted by the ability to detect errors. *American Journal of Surgery*. 2005 Apr;189(4):412–5.
83. Scott DJ, Bergen PC, Rege R V, Laycock R, Tesfay ST, Valentine RJ, et al. Laparoscopic Training on Bench Models : Better and More Cost Effective than Operating Room Experience?. *Journal of the American College of Surgeons*. 2000;7515(00):272–83.
84. Schueneman A, Pickleman J, Hesslein R, Freeark R. Neuropsychologic predictors of operative skill among general surgery residents. *Surgery*. 1984;96(2):288–95.
85. Winckel C, Reznick R, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *American Journal of Surgery*. 1994;167(4):423–7.
86. Jain N, Pietrobon R, Hocker S, Guller U, Shankar A, Higgins LD, et al. Volume and Outcomes. *Journal of Bone and Joint Surgery*. 2011;86(3):496–505.
87. Katz JN, Mahomed NN, Baron JA, Barrett JA, Fossel AH, Creel AH, et al. Association of hospital and surgeon procedure volume with patient-centered outcomes of total knee replacement in a population-based cohort of patients age 65 years and older. *Arthritis and Rheumatism*. 2007 Feb;56(2):568–74.
88. Archampong D, Borowski D, Lh I. Workload and surgeon’s specialty for outcome after colorectal cancer surgery (Review). *The Cochrane Collaboration*. 2012;(3).
89. Gruen R, Pitt V, Green S, Parkhill A, Campbell D, Jolley D. The Effect of Provider Case Volume on Cancer Mortality. *CA: A Cancer Journal for Clinicians*. 2009;59(3):192–211.
90. Ko CY, Chang JT, Chaudhry S, Kominski G. Are high-volume surgeons and hospitals the most important predictors of inhospital outcome for colon cancer resection? *Surgery*. 2002 Aug;132(2):268–73.
91. Brunner WC, Korndorffer JR, Sierra R, Massarweh NN, Dunne JB, Yau CL, et al. Laparoscopic virtual reality training: are 30 repetitions enough? *The Journal of Surgical Research*. 2004 Dec;122(2):150–6.
92. 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.
93. Alvand a, Auplish S, Khan T, Gill HS, Rees JL. Identifying orthopaedic surgeons of the future: the inability of some medical students to achieve competence in basic arthroscopic tasks despite training: a randomised study. *The Journal of Bone and Joint Surgery British Volume*. 2011 Dec;93(12):1586–91.

94. Pollard TCB, Tr F, Khan T, Price AJ, Gill HS, Glyn-Jones S, et al. Simulated hip arthroscopy skills: learning curves with the lateral and supine patient positions. *Journal of Bone and Joint Surgery*. 2012;68(1):1–10.
95. Howells NR, Auplish S, Hand GC, Gill HS, Carr AJ, Rees JL. Retention of arthroscopic shoulder skills learned with use of a simulator. Demonstration of a learning curve and loss of performance level after a time delay. *The Journal of Bone and Joint Surgery American Volume*. 2009 May;91(5):1207–13.
96. Howells NR, Brinsden MD, Gill RS, Carr AJ, Rees JL. Motion analysis: a validated method for showing skill levels in arthroscopy. *Arthroscopy*. 2008 Mar;24(3):335–42.
97. Xeroulis G, Dubrowski A, Leslie K. Simulation in laparoscopic surgery: a concurrent validity study for FLS. *Surgical Endoscopy*. 2009 Jan;23(1):161–5.
98. 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 Nov;193(5):479–85.
99. Grober ED, Roberts M, Shin E-J, Mahdi M, Bacal V. Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence. *American Journal of Surgery*. Elsevier Inc.; 2010 Jan;199(1):81–5.
100. Hayter M a, 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 Anaesthesia*. 2009 Jun;56(6):419–26.
101. 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.
102. TraumaMan Surgical Simulator. Simulab Corporate Website. 2013. Available from: <http://www.simulab.com/home-traumaman>
103. Wilson M, Middlebrook A, Sutton C, Stone R, McCloy R. MIST VR: a virtual reality trainer for laparoscopic surgery assesses performance. *Annals of the Royal College of Surgeons of England*. 1997;79(6):403–4.
104. Walsh C, Sherlock M, Ling S, Carnahan H. Virtual reality simulation training for health professions trainees in gastrointestinal endoscopy (Review). *The Cochrane Collaboration*. 2012;(6).
105. Gala R, Orejuela F, Gerten K, Lockrow E, Kilpatrick C, Chohan L, et al. Effect of Validated Skills Simulation on Operating Room Performance in Obstetrics and Gynecology Residents. *Obstetrics and Gynecology*. 2013;121(3):578–84.

106. Pokroy R, Du E, Alzaga A, Khodadadeh S, Steen D, Bachynski B, et al. Impact of simulator training on resident cataract surgery. *Graefe's Archive for Clinical and Experimental Ophthalmology*. 2013 Mar;251(3):777–81.
107. Zendejas B, Hernandez-Irizarry R, Farley DR. Does Simulation Training Improve Outcomes in Laparoscopic Procedures? *Advances in Surgery*. Elsevier Inc; 2012 Sep;46(1):61–71.
108. Fernandez GL, Page DW, Coe NP, Lee PC, Patterson L a, Skylizard L, et al. Boot cAMP: educational outcomes after 4 successive years of preparatory simulation-based training at onset of internship. *Journal of Surgical Education*. Elsevier Inc.; 2012;69(2):242–8.
109. Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills Learning curves and reliability measures. *Surical Endoscopy*. 2002;16(12):1746–52.
110. Gallagher A, Richie K, McClure N, McGuigan J. Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World Journal of Surgery*. 2001;25(11):1478–83.
111. Gallagher a. G, Lederman a. B, McGlade K, Satava RM, Smith CD. Discriminative validity of the Minimally Invasive Surgical Trainer in Virtual Reality (MIST-VR) using criteria levels based on expert performance. *Surgical Endoscopy*. 2004 Apr 1;18(4):660–5.
112. Matsumoto ED, Pace KT, D'a Honey RJ. Virtual reality ureteroscopy simulator as a valid tool for assessing endourological skills. *International Journal of Urology*. 2006 Jul 17;13(7):896–901.
113. Eriksen JR, Grantcharov T. Objective assessment of laparoscopic skills using a virtual reality stimulator. *Surgical Endoscopy*. 2005 Jul 28;19(9):1216–9.
114. Larsen CR, Grantcharov T, Aggarwal R, Tully a, Sørensen JL, Dalsgaard T, et al. Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surgical Endoscopy*. 2006 Sep;20(9):1460–6.
115. Schijven M, Jakimowicz J. Construct validity: experts and novices performing on the Xitact LS500 laparoscopy simulator. *Surgical Endoscopy*. 2003 May;17(5):803–10.
116. Schreuder HWR, Van Dongen KW, Roeleveld SJ, Schijven MP, Broeders I a MJ. Face and construct validity of virtual reality simulation of laparoscopic gynecologic surgery. *American Journal of Obstetrics and Gynecology*. Mosby, Inc.; 2009 May;200(5):540.e1–8.
117. Van Dongen KW, Tournoij E, Van der Zee DC, Schijven MP, Broeders IA. Construct validity of the LapSim: can the LapSim virtual reality simulator distinguish between novices and experts? *Surgical Endoscopy*. 2007 Aug;21(8):1413–7.

118. Tran LN, Gupta P, Poniatowski LH, Alanee S, Dall'era M a, Sweet RM. Validation study of a computer-based open surgical trainer: SimPraxis(®) simulation platform. *Advances in Medical Education and Practice*. 2013 Jan;4:23–30.
119. Beier F, Diederich S, Schmieder K, Manner R. NeuroSim - the prototype of a neurosurgical training simulator. *Studies in Health Technology and Informatics*. 2011;(163):51–6.
120. Kuroda Y, Hirai M, Nakao M, Sato T, Kuroda T, Nagase K, et al. Organ exclusion simulation with multi-finger haptic interaction for open surgery simulator. *Studies in Health Technology and Informatics*. 2007;(125):244–9.
121. Tsai MD, Hsieh MS, Jou SB. Virtual reality orthopedic surgery simulator. *Computers in Biology and Medicine*. 2001 Sep;31(5):333–51.
122. Sanders A, Luursema J, Warntjes P, Mastboom W, Geelkerken R, Klasse J, et al. Validation of open-surgery VR trainer. *Studies in Health Technology and Informatics*. 2006;(119):473–6.
123. Sorensen T, Stawiaski J, Mosegaard J. Virtual open heart surgery: obtaining models suitable for surgical simulation. *Studies in Health Technology and Informatics*. 2007;(125):445–7.
124. Botden SMBI, Buzink SN, Schijven MP, Jakimowicz JJ. Augmented versus virtual reality laparoscopic simulation: what is the difference? A comparison of the ProMIS augmented reality laparoscopic simulator versus LapSim virtual reality laparoscopic simulator. *World Journal of Surgery*. 2007 Apr;31(4):764–72.
125. MedSim Corporation. MedSim Advanced Ultrasound Simulation. 2013. Available from: <http://www.medsim.com/>
126. CAE Healthcare Corporation. CAE Healthcare - Product & Services. 2013. Available from: https://caehealthcare.com/home/eng/product_services/
127. Symbionix Corporation. Symbionix US Mentor. 2013. Available from: <http://symbionix.com/simulators/us-mentor/>
128. Sheehan F, Otto C, Freeman R. Echo simulator with novel training and competency testing tools. *Studies in Health Technology and Informatics*. 2013;184:397–403.
129. Burden C, Preshaw J, White P, Draycott TJ, Grant S, Fox R. Validation of virtual reality simulation for obstetric ultrasonography: a prospective cross-sectional study. *Simulation in Healthcare*. 2012 Oct;7(5):269–73.
130. Knudson MM, Sisley AC. Training residents using simulation technology: experience with ultrasound for trauma. *The Journal of Trauma*. 2000 Apr;48(4):659–65.

131. Bann S, Kwok K-F, Lo C-Y, Darzi A, Wong J. Objective assessment of technical skills of surgical trainees in Hong Kong. *The British Journal of Surgery*. 2003 Oct;90(10):1294–9.
132. Bann S, Datta V, Khan M, Darzi A. The surgical error examination is a novel method for objective technical knowledge assessment. *The American Journal of Surgery*. 2003 Jun;185(6):507–11.
133. Sites B, Gallagher J, Cravero J, Lundberg J, Blike G. The learning curve associated with a simulated ultrasound-guided interventional task by inexperienced anesthesia residents. *Regional Anesthesia and Pain Medicine*. 2004 Dec;29(6):544–8.
134. Beese R, Lowe S. The use of turkey breast and stuffed olives as a soft tissue model for the teaching and practice of ultrasound guided interventional procedures. *European Journal of Ultrasound*. 1998;7(1):12.
135. Sultan SF, Iohom G, Saunders J, Shorten G. A clinical assessment tool for ultrasound-guided axillary brachial plexus block. *Acta Anaesthesiologica Scandinavica*. 2012 May;56(5):616–23.
136. Sarker SK, Chang A, Vincent C, Darzi SAW. Development of assessing generic and specific technical skills in laparoscopic surgery. *American Journal of Surgery*. 2006 Feb;191(2):238–44.
137. Datta V, Bann S, Beard J, Mandalia M, Darzi A. Comparison of bench test evaluations of surgical skill with live operating performance assessments. *Journal of the American College of Surgeons*. 2004;199(4):603–6.
138. Streiner D. Global Rating Scales in: *Assessing Clinical Competence*. Assessing Clinical Competence. 1985. p. 119–40.
139. Birnbach DJ, Santos AC, Bourlier R a, Meadows WE, Datta S, Stein DJ, et al. The effectiveness of video technology as an adjunct to teach and evaluate epidural anesthesia performance skills. *Anesthesiology*. 2002 Jan;96(1):5–9.
140. Sultan S, Iohom G, Saunders J, Shorten G. A Clinical Assessment Tool for Ultrasound-Guided Axillary Brachial Plexus Block. *Acta Anaesthesiologica Scand*. 2012;56(5):616–23.
141. Vogt V. Is a resident's score on a videotaped objective structured assessment of technical skills affected by revealing the resident's identity? *American Journal of Obstetrics and Gynecology*. 2003 Sep;189(3):688–91.
142. Ascension Technology Corporation. 3D Guidance trakSTAR - Ascension Technology Corporation. 2013. Available from: <http://www.ascension-tech.com/medical/trakSTAR.php>

143. 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.
144. 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. *Annals of Surgery*. 2004 Aug;240(2):374–81.
145. 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 Mar;167(3):1243–7.
146. Hennepin County Medical Center. FAST Exam Hennepin County. Online. 2008. Available from: <http://vimeo.com/1044031>
147. FujiFilm Sonosite Incorporated. SonoSite M Turbo. Online. 2013. Available from: <http://www.sonosite.com/products/m-turbo>
148. Davison a. C, Hinkley D V. *Bootstrap methods and their application*. Cambridge: Cambridge University Press; 1997.
149. Fink a, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *American Journal of Public Health*. 1984 Sep;74(9):979–83.
150. Williams PL, Webb C. The Delphi technique: a methodological discussion. *Journal of Advanced Nursing*. 1994 Jan;19(1):180–6.
151. Cheung JJH, Chen EW, Darani R, McCartney CJL, Dubrowski A, Awad IT. The creation of an objective assessment tool for ultrasound-guided regional anesthesia using the Delphi method. *Regional Anesthesia and Pain Medicine*. 2012;37(3):329–33.
152. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Family Medicine*. 2005 May;37(5):360–3.
153. Dosis A, Aggarwal R, Bello F. Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Archives of Surgery*. 2005;140.
154. Dosis A, Bello F, Moorthy K. 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.
155. Sivaprakasam J, Purva M. CUSUM analysis to assess competence: what failure rate is acceptable? *The Clinical Teacher*. 2010;257–61.
156. Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. Value of an objective assessment tool in the operating room. *Canadian Journal of Surgery*. 2011 Apr;54(2):116–22.

157. Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. *Journal of the American College of Surgeons*. Elsevier Inc.; 2009 Sep;209(3):364–370.e2.

Appendix A

FAST Image Acquisition Evaluation Checklist

Hepatorenal Space

<input type="checkbox"/> Orients image with the liver to the left and kidney to the right
<input type="checkbox"/> Adjusts depth so the image ends just below the kidney
<input type="checkbox"/> Sets gain appropriately
<input type="checkbox"/> Visualizes the interface between the liver and kidney clearly
<input type="checkbox"/> Visualizes the interface between the liver and kidney in entirety by sweeping through the entire kidney
<input type="checkbox"/> Visualizes the caudal tip of the liver clearly

Splenorenal Space

<input type="checkbox"/> Orients image with the spleen to the left and kidney to the right
<input type="checkbox"/> Adjusts depth so the image ends just below the kidney
<input type="checkbox"/> Sets gain appropriately
<input type="checkbox"/> Visualizes the interface between the spleen and kidney clearly
<input type="checkbox"/> Visualizes the interface between the spleen and kidney in entirety by sweeping through the entire kidney
<input type="checkbox"/> Clearly visualizes space between diaphragm and spleen

Pelvis

<input type="checkbox"/> Adjusts depth so the image ends 4-5 cm below the bladder
<input type="checkbox"/> Sets gain appropriately such that the urine in the bladder appears black
<input type="checkbox"/> Visualizes the bladder in longitudinal section
<input type="checkbox"/> Visualizes the bladder in entirety in longitudinal section by scrolling through the entire bladder
<input type="checkbox"/> Visualizes the bladder in transverse section
<input type="checkbox"/> Visualizes the bladder in entirety in transverse section by sweeping through the entire bladder

Pericardium

<input type="checkbox"/> Orients image such that the apex of the ventricles point towards the right of the image
<input type="checkbox"/> Adjusts depth so the image ends just past the deepest layer of pericardium (relative to ultrasound probe)
<input type="checkbox"/> Sets gain appropriately such that the blood in the ventricles appears black
<input type="checkbox"/> Optimizes view of pericardium using adjuncts as necessary (ex. Parasternal view, breath holding)
<input type="checkbox"/> Visualizes both the anterior and posterior pericardium
<input type="checkbox"/> Visualizes the pericardium in its entirety by sweeping through the entire heart

Appendix B

FAST Representative Image Evaluation Checklist

Hepatorenal Space

<input type="checkbox"/> The image is oriented with the liver to the left and kidney to the right
<input type="checkbox"/> Depth is appropriately adjusted so the image ends just below the kidney
<input type="checkbox"/> Gain is set appropriately
<input type="checkbox"/> The interface between the liver and kidney is clearly seen
<input type="checkbox"/> The caudal tip of the liver is clearly seen

Splenorenal Space

<input type="checkbox"/> The image is oriented with the spleen to the left and kidney to the right
<input type="checkbox"/> Depth is appropriately adjusted so the image ends just below the kidney
<input type="checkbox"/> Gain is set appropriately
<input type="checkbox"/> The interface between the spleen and kidney is clearly seen
<input type="checkbox"/> The caudal tip of the spleen is clearly visualized

Pelvis

<input type="checkbox"/> Depth is appropriately adjusted so the image ends 4-5 cm below the bladder
<input type="checkbox"/> Gain is set appropriately such that the urine in the bladder appears black

Pericardium

<input type="checkbox"/> The image is oriented such that the apex of the ventricles point towards the right of the image
<input type="checkbox"/> Depth is appropriately adjusted so the image ends just past the superior pericardium
<input type="checkbox"/> Gain is set appropriately such that the blood in the ventricles appears black
<input type="checkbox"/> Both the inferior and superior pericardium are visualized

Appendix C

Global Rating Scale of FAST Image Acquisition

1. Skin Contact				
1	2	3	4	5
<i>Consistently uses insufficient amounts of gel or achieves inadequate skin contact</i>		Uses <i>appropriate</i> amounts of gel and achieves <i>adequate</i> skin contact <i>most</i> of the time		<i>Consistently uses appropriate amounts of gel and achieves adequate skin contact</i>

2. Image Adjustment				
1	2	3	4	5
<i>Inappropriately sets gain or depth</i>		Adjusts gain and depth <i>appropriately</i> but <i>occasionally</i> requires repeat adjustment throughout the procedure		Adjusts gain and depth to an <i>appropriate</i> level only once at the <i>beginning</i> of each section

3. Initial Probe Placement				
1	2	3	4	5
<i>Frequently readjusts probe position on the skin or obtains inadequate views</i>		<i>Correctly</i> places the probe to obtain <i>adequate</i> views but <i>occasionally</i> requires readjustment		<i>Correctly</i> places the probe to obtain <i>adequate</i> views on <i>1st</i> attempt with <i>minimal</i> readjustment

4. Image Sweeping				
1	2	3	4	5
After establishing probe position <i>continues</i> to reposition in a <i>staccato</i> manner		After establishing probe position has <i>mostly smooth</i> image sweeping but makes occasional <i>staccato</i> movements		After establishing probe position makes <i>subtle</i> probe movements with <i>smooth</i> sweeping

5. Sonographer Positioning and Probe Handling				
1	2	3	4	5
Repeatedly assumes an <i>awkward</i> body position or holds the probe in an <i>awkward</i> or <i>inappropriate</i> manner		Occasionally assumes an <i>awkward</i> body position or holds the probe in an <i>inappropriate</i> manner		Assumes a <i>comfortable</i> body position and holds the probe in an <i>appropriate</i> manner

6. Time of Exam				
1	2	3	4	5
Takes an <i>excessively</i> long time to complete the exam		Completes the exam in an <i>average</i> amount of time		<i>Rapidly</i> completes the exam <i>with</i> acceptable performance

7. Flow of Procedure				
1	2	3	4	5
Is <i>consistently unorganized</i> with <i>frequent</i> jumps between anatomic regions		<i>Mostly</i> organized but jumps <i>occasionally</i> between anatomic regions		Completes the procedure by moving <i>smoothly</i> from region to region

8. Autonomy				
1	2	3	4	5
Unable to complete exam without <i>significant</i> guidance		Able to complete task correctly with <i>moderate</i> guidance		Able to complete task <i>independently</i> without prompting

9. 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

Appendix D



Faculty of Medicine
Department of Surgery

3rd Floor Z Block, 409 Tache Avenue
Winnipeg, MB, Canada R2H 2A6
Ph (204) 237-2574
Fax (204) 237-3429

Title of Study: “Standardized Assessment of Ultrasound Image Acquisition”

**Principal Investigator: Dr. Lawrence Gillman, St. Boniface Hospital, 409 Tache Avenue
Z-Block 3rd Floor, (204) 258-1408**

You are being asked to participate in a research study. Please take your time to review this consent form and discuss any questions you may have with the study staff. You may take your time to make your decision about participating in this study and you may discuss it with your friends, family or (if applicable) your doctor before you make your decision. This consent form may contain words that you do not understand. Please ask the study staff to explain any words or information that you do not clearly understand.

Purpose of Study

The purpose of this study is to develop an assessment tool for the Focused Assessment with Sonography (FAST) exam. The information from this study can help us give feedback to students during courses and curricula for FAST training and help us define who is competent to perform a FAST exam.

A total of 24 participants will participate in this study

Study procedures

You will watch a standardized video outlining the FAST procedure. You will then be asked to perform a FAST exam on a standardized patient while being video-recorded and assessed.

You can stop participating at any time. However, if you decide to stop participating in the study, we encourage you to talk to the study staff first.

Risks and Discomforts

There are no known harms associated with study participation. You may experience uneasiness when having your ultrasound skills assessed. We therefore emphasize that you will remain anonymous and may also withdraw from the study at any time.

Benefits

You will gain training in the FAST exam procedure.

Costs

All the procedures, which will be performed as part of this study, are provided at no cost to you.

Payment for participation

You will receive no payment or reimbursement for any expenses related to taking part in this study.

Confidentiality

Information gathered in this research study may be published or presented in public forums, however your name and other identifying information will not be used or revealed. All study related documents and video files will bear only your assigned study number and you will not be identifiable in the video. Despite efforts to keep your personal information confidential, absolute confidentiality cannot be guaranteed. Your personal information may be disclosed if required by law.

The University of Manitoba Health Research Ethics Board and St. Boniface Hospital may review records related to the study for quality assurance purposes.

Voluntary Participation/Withdrawal from the Study

Your decision to take part in this study is voluntary. You may refuse to participate or you may withdraw from the study at any time. Your decision not to participate or to withdraw from the study will not affect your evaluation or your relationship with your colleagues or supervisors at this centre.

Questions

You are free to ask any questions that you may have about your treatment and your rights as a research participant. If any questions come up during or after the study or if you have a research-related injury, contact the study doctor and the study staff: Dr. Lawrence Gillman at (204) 258-1408.

For questions about your rights as a research participant, you may contact The University of Manitoba, Bannatyne Campus Research Ethics Board Office at (204) 789-3389

Do not sign this consent form unless you have had a chance to ask questions and have received satisfactory answers to all of your questions.

Statement of Consent

I have read this consent form. I have had the opportunity to discuss this research study with Dr. Lawrence Gillman or his study staff. I have had my questions answered by them in language I understand. The risks and benefits have been explained to me. I believe that I have not been unduly influenced by any study team member to participate in the research study by any statements or implied statements. Any relationship (such as employer, supervisor or family member) I may have with the study team has not affected my decision to participate. I understand that I will be given a copy of this consent form after signing it. I understand that my participation in this study is voluntary and that I may choose to withdraw at any time. I freely agree to participate in this research study.

I understand that information regarding my personal identity will be kept confidential, but that confidentiality is not guaranteed.

By signing this consent form, I have not waived any of the legal rights that I have as a participant in a research study.

Participant signature _____ Date _____

Participant printed name: _____

I, the undersigned, have fully explained the relevant details of this research study to the participant named above and believe that the participant has understood and has knowingly given their consent

Printed Name: _____ Date _____

(day/month/year)

Signature: _____

“Role in the study: _____

Relationship (if any) to study team members:_____

Appendix E

FAST Study Questionnaire

Novice 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