

**Validation of a Simulation-Based Tool
For Formative Assessment of Echocardiography Skill Competence**

BY

SHERRYN PRIYA RAMBIHAR
B.Sc., University of Toronto, 2001
M.D., Western University, 2005

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Health Professions Education
In the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:

Ara Tekian, Chair and Advisor
Ryan Brydges, University of Toronto
Matthew Lineberry, University of Kansas
Yoon Soo Park

This thesis is dedicated to: Conor, Reid and Rory.

“Time passes and it tells us what we're left with, we become the things we do.”

☛ *Third Eye Blind, Staring Down the Sun*

ACKNOWLEDGEMENTS

I would like to thank the following individuals and institutions for their support and assistance.

Thesis committee

- Dr. Ara Tekian, Chair and Advisor, Department of Medical Education, University of Illinois at Chicago
- Dr. Ryan Brydges, The Wilson Centre and Li Ka Shing Knowledge Institute, University of Toronto
- Dr. Matthew Lineberry, Zamierowski Institute for Experiential Learning, University of Kansas Medical Center
- Dr. Yoon Soo Park, Department of Medical Education, University of Illinois at Chicago

Research Assistants and Additional Support

- Dr. Jeremy Edwards, Division of Cardiology, University of Toronto
- Dr. Kristina Khanduja, Department of Anesthesia, University of Toronto
- Ella Kisselman, Sonographer, Women's College Hospital
- Dr. Gera Kisselman, Internal Medicine Resident, University of Toronto
- Dr. Gillian Nesbitt, Division of Cardiology, University of Toronto
- University of Illinois at Chicago Masters of Health Professions Education Program
- Wei Wu, Statistician, Women's College Research Institute

Financial support

- Dr. Paul Dorian, Departmental Director of Cardiology, University of Toronto
- Cardiology Academic Fund, Women's College Hospital
- Dr. Rodrigo Cavalcanti, The HoPingKong Centre for Excellence in Clinical Practice, University of Toronto

SR

TABLE OF CONTENTS

<u>CHAPTER</u>	<u>PAGE</u>
<u>1: INTRODUCTION</u>	1
<u>2: CONCEPTUAL FRAMEWORKS</u>	8
2.1 Validity Theory	8
2.2 Early Validity Frameworks	8
2.3 Messick's Unified Framework for Validity	9
2.4 Challenges of Messick's Unified Framework	10
2.5 Kane's Argument-based Approach to Validity	11
<u>3: REVIEW OF RELEVANT LITERATURE</u>	14
3.1 Applications of Kane's Framework in Medical Education	14
3.2 Applications of Kane's Framework to Echocardiography Competency Assessment	19
3.3 Applications of Kane's Framework in our Context	27
<u>4: RESEARCH QUESTION AND HYPOTHESES</u>	29
4.1 Research Question	29
4.2 Interpretation-use Argument	29
4.3 Hypotheses	29
<u>5: METHODS</u>	33
5.1 Sample Size and Participants	33
5.2 Setting	33
5.3 Design	34
5.4 Materials	34
5.5 Procedure	41
5.6 Rater Training	44
5.7 Analysis	45
5.8 Reflexivity	47
<u>6: RESULTS</u>	48
6.1 Demographics	48
6.2 Evidence for Implications	48
6.3 Evidence for Scoring	65
6.4 Evidence for Extrapolation	68
<u>7: DISCUSSION</u>	72
7.1 Validity Argument: Judgment on our Interpretation-use Argument	72
7.1.1 The Implications Inference	72
7.1.2 The Scoring Inference	76
7.1.3 The Extrapolation Inference	77

TABLE OF CONTENTS (CONTINUED)

<u>CHAPTER</u>	<u>PAGE</u>
<u>7: DISCUSSION</u>	72
7.1.4 The Generalization Inference	80
7.1.5 Other Factors Impacting the Validity Argument	80
7.1.6 Summary Judgment	81
7.2 Impact on Future Research	81
7.3 Limitations	83
<u>8: CONCLUSIONS</u>	87
<u>CITED LITERATURE</u>	88
<u>APPENDICES</u>	97
Appendix A	97
Appendix B	99
Appendix C	100
Appendix D	102
Appendix E	103
Appendix F	104
Appendix G	123
Appendix H	124
Appendix I.....	125
Appendix J	128
Appendix K	130
<u>VITA</u>	131

LIST OF TABLES

<u>TABLE</u>	<u>PAGE</u>
I. SUMMARY OF INFERENCE CATEGORIES, APPRAISAL OF EXISTING EVIDENCE, RATIONALE FOR PRIORITIZATION, CLAIMS AND HYPOTHESES AND ASSOCIATED METHODS AND ANALYSES	31
II. PARTICIPANT DEMOGRAPHICS	48
III. LIST OF INITIAL CODES AND FINAL THEMES	49
IV. TEST ITEM DESCRIPTIVE STATISTICS	66
V. EXPLORATORY ROTATED FACTOR PATTERN ANALYSIS	67
VI. CORRELATION BETWEEN ECAT SCORE AND OTHER VARIABLES	70
VII. PAIRWISE COMPARISONS BETWEEN PARTICIPANT GROUPS	71

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1.	ECAT study protocol	41
2.	Association between ECAT score and diagnostic quality assessment	68
3.	Correlation between ECAT score and summative exam score	69
4.	ECAT score according to level of training	70

LIST OF ABBREVIATIONS

ACC	American College of Cardiology
ACGME	American College of Graduate Medical Education
ASE	American Society of Echocardiography
CBME	Competency-based Medical Education
CL	Task-specific Checklist
CoCATS	Core Cardiovascular Training Statement
D-study	Decision Study
ECAT	Echocardiography Competence Assessment Tool
FATE	Focused Anesthesia Transthoracic Echocardiogram
G-study	Generalizability study
GMA	General Mental Ability
GRS	Global Rating Scale
MCAT	Medical College Admissions Test
mCEX	Mini-Clinical Evaluation Exercise
OSATS	Objective Structured Assessment of Technical Skills
OSAUS	Observed structured Assessment of Ultrasound Skills
OSCE	Observed Structured Clinical Examination
PGY	Postgraduate Year
TEE	Transesophageal Echocardiography

SUMMARY

Competency-based medical education (CBME) frameworks suggest including robust formative assessments throughout residency training to support learners' skills development and self-regulated learning. We developed a simulation-based echocardiography competence assessment tool (ECAT) for formative assessment of basic echocardiography skills in a cardiology training program, and evaluated its validity evidence as outlined in Kane's argument-based approach to validation [1].

In two study phases, we modified the ECAT for formative assessment then assessed residents. The assessment involved participant orientation to an echocardiography simulator, hands-on practice time, performance assessment using the ECAT, and provision of written feedback. Participants were interviewed two weeks later, and raters interviewed at study completion. We used the following analyses to evaluate our validity argument:

- 1) Implications – We sought evidence that participants and raters perceived that ECAT testing improved echocardiography skills and facilitated provision of feedback. We analyzed interview data from participants and raters using a simple content analysis framework to understand the impact of our assessment.
- 2) Scoring – We sought evidence that ECAT scoring procedures represented observed echocardiography skills. We designed the ECAT scoring rubric using professional standards as a blueprint, analyzed internal consistency between dimensions on the ECAT using factor analysis, computed rater reliability, and explored participant and rater narratives.
- 3) Extrapolation – We sought evidence that the ECAT score translated to clinical skill performance. We examined the association between ECAT scores and summative testing, global diagnostic scan quality, and level of training, to clarify if observed relationships met our expectations.

Our analysis of the interview data revealed three themes: i) feedback stimulated change, ii) how feedback was delivered impacted participants' perceived learning, and iii) assessment credibility influenced participants' reception of feedback. For scoring evidence, inter-rater reliability was ICC=0.913, and an exploratory factor analysis demonstrated a two-factor model, for which Cronbach's alphas were 0.96 and 0.87 respectively. For extrapolation evidence, ECAT scores correlated with summative exam scores ($r=0.66$, $p=0.02$), and were positively associated with level of training ($p=0.006$), previous echocardiography experience ($p=0.01$), and diagnostic scan quality ($p=0.0006$).

This study is the first to systematically generate validity evidence for a tool to assess basic echocardiography skills for formative simulation-based assessment of trainees. We found the ECAT to be a valued tool which was thought to be meaningful to both participants and raters in helping with goal setting and provision of feedback, demonstrated good inter-rater reliability, and correlated well with level of expertise and a summative assessment. The process of evaluating the appropriateness of the interpretations, uses and decisions stemming from assessment results will assist in future development of instruments for CBME.

1: INTRODUCTION

Competency-based medical education requires robust assessments of competence

Competency-based medical education (CBME) is an important part of the transformation of medical education for the 21st century [2]. CBME frameworks propose that the focus of education and evaluation shifts from historically educator-centered, time-based approaches to learner-centered mastery approaches [3-6]. Educators designing CBME curricula ask, “what abilities are needed of medical graduates,” identify multiple competencies for domains of abilities at different stages of medical training, and select instructional methods and assessment tools to facilitate development of these abilities over time [3,7,8]. An effective assessment strategy to support development of competence is frequent assessment and progress testing, in a model described as “assessment for learning” [9]. As CBME becomes widely operationalized under the ACGME Milestone system, CanMEDS 2015 educational framework, and others [10-15], it is vitally important that educators collect validity evidence to evaluate the quality of the many tools they use to form judgments of learners’ competence, and their progression and skills development [1].

Psychometric challenges in competency-based assessment

Research in the field of competency-based assessment has revealed several psychometric challenges. Unlike simple skills which follow a predictable learning growth curve, complex skills may develop in a dynamic, non-linear fashion, with multiple time points where competence is achieved or recedes, before ultimate attainment of skill level consistent with independent

practice [3,16,17]. The optimal number of observations to capture a comprehensive view of any individual's competence is thus unclear, and potentially unrealistic [17]. An assessment of competency for general CanMEDS roles, for instance, demonstrated poor rater reliability, and would require an impractically large number of ratings to yield acceptable results [17]. Raters using scales to assess broad competencies have poor agreement and an inability to discern performance for these categories [16-18]. These difficulties highlight that the use of existing tools to assess broadly-defined competencies has led to challenges identifying precisely when and whether individuals have achieved a prescribed level of competence [16].

Specialty-specific cardiology competencies defined incompletely

Defining competencies more specifically may improve overall assessment, and organizations continue to call for specialty-specific CBME approaches. Cardiology represents a useful example for studying assessment and CBME because mastering the skills required to practice cardiology in three years of training is becoming more difficult, due to increased cognitive and procedural skills required, and competing forces including post-call days and work-hour restrictions limiting time for teaching and acquisition of skills [19-22]. American cardiology programs assess trainees according to 142 core competencies defined by the ACGME Internal Medicine Specialty Milestones Project [10-12]; however, there is a perception from cardiology educators that this framework does not adequately reflect high-priority cardiology core competencies [23].

Cardiology training is unique amongst subspecialties for skills where the learner is both technician and interpreter, and must integrate motor skills, judgment and medical knowledge in real-time to make a diagnosis and produce a complete repository of images to be used clinically

[23,24]. Thus, while specialty-specific competencies are desirable, practitioners defining these competencies must capture essential skills without resulting in impractical, endless lists [3].

Echocardiography is an essential cardiology milestone

Performing and interpreting an echocardiogram is an example of an essential cardiology milestone. Echocardiography is a complex skill that uses sound waves to create moving pictures of the heart. Echocardiography requires integrating multiple cognitive domains: knowledge of acoustic physics, hand-eye coordination to generate diagnostic images from positioning a transducer at various sites on the chest wall, visual-spatial-conceptual allocation of acquired two-dimensional images to the three-dimensional moving heart, and interpretation of images in the context of cardiac hemodynamics in healthy and pathologic conditions [25].

Echocardiography is the most widely used and readily available imaging technique in cardiology, often used in emergencies to diagnose and manage conditions at the bedside [24].

Echocardiography is a skill learners develop initially and hone through ongoing experience throughout residency. The process of how learning occurs in echocardiography has not been well studied. However, echocardiography seems well suited as a target for formative CBME approaches, as the skills develop slowly and, anecdotally, trainees report a need for ongoing feedback.

Challenges in how echocardiography is taught and assessed

The advent of clinical echocardiography in the 1970s increased interest in how cardiologists learn about echocardiography [26], but there remains a paucity of research in effective

instructional methods to support development of competence. Educators teach residents to perform and interpret echocardiograms through a “master-apprentice” model, over six rotations in a three-year residency [27,28]. Well-defined standards outline knowledge, technical skills, evaluation skills, and attitudes expected by completion of residency, and competence has been historically assumed based on number of scans performed at the end of the training program, duration of training, and an In-Training Evaluation Report [27-29]. However, these metrics used to infer competence in graduating residency have not correlated well with proficiency in previous studies [29, 30]. There is a mismatch between what trainees perceive as necessary to achieve competence, compared to the recommended volume targets for competency, and most cardiology training programs are unaware of recommended volumes, much less how best to design rotations to promote development of procedural and interpretive skills [19]. Additional realities paving the way for CBME in echocardiography include the clinical laboratory focus on throughput, with inconsistent educator availability for coaching, and the recognition that learners have no objective way to benchmark skills development during training, leaving them without areas to focus on for self-regulated learning [25].

Proposing a shift to competency-based echocardiography education

Frank [3] proposes a six-step process to plan CBME curricula: “identify the abilities needed of graduates, explicitly define the required competencies and their components, define milestones along a development path for the competencies, select educational activities, experiences and instructional methods, select assessment tools to measure progress along the milestones and design an outcomes program evaluation.” As an initial step in echocardiography, we attempted

to define the knowledge, technical and evaluative skills and attitudes of novices in the echocardiography lab [25]. That study revealed that technical skills required of novice trainees were poorly defined, limiting instructors' abilities to provide formative feedback tailored to trainee needs [25].

Recently, a task force of clinical cardiology and medical education experts revised echocardiography training standards using a CBME framework [21, [Appendix A](#)]. This comprehensive roadmap creates a matrix with timelines in the three years in which trainees are expected to acquire specific competencies clustered by ACGME competency categories (knowledge, patient care and procedural skills, system-based practice, practice-based learning, professionalism and interpersonal skills). The map also includes training requirements of the American Society of Echocardiography (ASE) for level 1 expertise (early learners), level 2 expertise (independent practitioners at the end of residency), and level 3 expertise (advanced expertise requiring post-residency training). Each ACGME domain lists suggested tools for assessing formative and summative competence, including conference presentation, direct observation, in-training examination, logbook, simulation, multi-source evaluation, reflection, self-assessment, and National Board of Echocardiography Exam. However, details are not provided to facilitate implementation, and the steps to ensure progress and competence are left to the discretion of program directors and Clinical Competence Committees [24,31].

Simulation to facilitate competency-based assessment

Simulation is an educational method which “attempts to present [education and] evaluation problems authentically... the trainee is required to respond to the problems as he or she would under natural circumstances. Frequently, the trainee receives performance feedback” [32]. Some propose that simulation facilitates formative and summative assessment, and could be useful for echocardiography competence assessments together with other tools outlined in new CBME echocardiography training standards [24, 27].

In contrast to the high-stress, variable clinical learning environment, a simulation setting is thought to afford learners the opportunity to develop skills proficiency through deliberate practice in a reproducible, standardized environment according to their time and individual learning needs [32-34]. A commercial echocardiography simulator exists, and provides haptic feedback through the sensory input of holding a probe to generate simulated images, with dynamic real-time guidance of the anatomy encountered and demonstration of how plane changes alter visualized structures [35, 36]. However, there are several caveats to this technology: computer-generated images may be too synthetic; the mannequin cannot be positioned or coached like a patient in maneuvers to optimize image quality; computerized patient difficulty modes appear artificial; transducer locations where images are found are present or absent in a binary fashion, rather than the intermediate zones where images are partially found in patients; a standard echocardiography window is missing; and hovering the probe over the mannequin without making contact can generate images. Initial research on echocardiography simulation from anesthesia and critical care perspectives has provided a

rationale for its use by comparisons with standard echocardiography training [37-39]. These studies are observational, and focused on efficacy of an intervention; thus they have limited application to research and practice. No studies to date have clarified how and why simulation in echocardiography may be useful using conceptual frameworks, in order to advance our understanding beyond prior descriptive studies.

Synthesis

Cardiology may be representative of training programs experiencing challenges in how to support skills development, and there is little research on effective instructional methods to support development of competence. Echocardiography is a suitable clinical skill for studying formative assessment, given skill development occurs throughout training and there is a reported need for feedback. The choice of simulation as modality to assess skills permits study under standardized conditions with a focus on feedback. We propose that research is needed to respond to the needs outlined in new echocardiography professional training standards [21], and to establish evidence that informs the implementation and future investigation of CBME. This thesis seeks to clarify the use of the echocardiography competence assessment tool (ECAT) for formative assessment of cardiology trainees' simulation-based basic echocardiography skills, using conceptual frameworks from contemporary validity theory [1].

2: CONCEPTUAL FRAMEWORKS

2.1 Validity Theory

Validity theory underpins the collection, interpretation, and use of evidence for performance assessment [1,40,41]. Validity evidence imbues the observed behaviors we measure in performance assessment with meaning, and permits us to draw specific conclusions in specific contexts for specific populations [42]. Messick defines validity as “the degree to which evidence and theory support [or refute] the interpretations of test scores for proposed uses” [43]. Thus, validity is a property of score interpretations and uses, rather than a property of the tool itself [44]. One direction in which validity theory has evolved over the past century is toward an argument-based framework with “interpretation-use arguments” guiding evidence analysis [1]. Validity theory is the basis for ideas explored in this thesis.

2.2 Early Validity Frameworks

The historical precedent for educational validity theory from 1920-1950 was the correlation between test scores and the ‘true’ criterion score from actual task performance, defined by Cureton as “criterion validity” [45]. This approach was effective when a good reference standard was available, however with less tangible attributes, selection of adequate alternative criteria was controversial and fraught with value judgments [40,46]. Alternative approaches for validation of constructs were sought [1].

2.3 Messick's Unified Framework for Validity

Messick's unified framework for validity, the current field standard for the American Educational Research Association [47], shifted the emphasis from validation of the test to "the development and validation of a proposed interpretation of test scores" [1]. Five distinct sources of validity evidence were unified under the umbrella of "construct validity", "a measure of an attribute for which an adequate criterion could not be defined" where validity is assessed by measuring observable attributes with theorized relationships [40]. The five types of evidence were: content, response process, internal structure, relationship to other variables, and implications; criterion validity was relegated to an ancillary methodology [42,43,48]. Different types of assessments require different types of evidence, which when taken together, support or refute specific hypothetical interpretations of the assessment data [42,48].

Content evidence is found in an assessment where sampled performance estimates overall skill [40,48]. This is achieved by having a large, representative sample of observations agreed upon by content experts, fair performance evaluation, and blueprinting to relate test content to learning objectives [1,42,48]. Some critics feel the selection of test tasks can reflect test developers' confirmation bias, and that content evidence may play a limited and more basic role in acquisition of validity evidence compared to the other types of evidence [43].

Response process evidence is evidence of how well the documented response on assessment testing reflects the observed performance. High quality evidence is found when rater error and cheating is minimized [48].

Internal structure evidence is derived from psychometric analyses of individual assessment items with each other and with the construct, such as reliability and reproducibility of data and scores, correlations between items, factor analysis, and interactions [48].

Relationship to other variables evidence arises from correlations between assessment scores and other measures with pre-specified theoretical relationships such as different instruments or comparing learner groups expected to differ. Results may not be generalizable, and interpretation of these correlations must consider individual study design and limitations [48].

Consequences validity evidence may be the most important type of evidence, as it reflects “the impact, beneficial or harmful, and unintended or intended, of assessment, and the decisions and actions that result, and factors that directly influence the rigor of such decisions” [4].

Assessments have impact on learners, teachers, people, and systems they influence [50]. Two dimensions of implications evidence can be considered in terms of their impact: assessments themselves, and the consequences of the resulting scores [41].

2.4 Challenges of Messick’s Unified Framework

Messick’s unified framework reflected a fundamental change in how validity was considered.

Validation came to be understood as a process to evaluate the appropriateness of interpretations of test scores through critical analysis of various types of evidence, rather than an endpoint [42, 44]. However, a key critique of the Messick approach was while this unified

framework yielded multiple sources of evidence and could be applied widely to different types of

assessments, it did not offer any guidance on how to prioritize between evidence required for different types and purposes of assessments [40,50].

2.5 Kane’s Argument-based Approach to Validity

A complementary argument-based approach to validity was first proposed by Cronbach [51] and Messick [43], and elaborated more recently by Kane [1,40,41]. The Kane framework (Appendix B, C), takes the evidence identified by unified validity theory and goes one step further, starting with a structured hypothesis for the interpretation of assessment scores, and critically curating the selection, interpretation and presentation of evidence to make a judgment on the soundness of test interpretation [52]. Just as a hypothesis cannot be proven, validity can never be proven, but accumulating evidence can support or refute the validity argument [44]. Expanding on Messick’s framework, educators may find Kane’s framework “accessible and applicable to a wide range of assessment tools and activities” [40].

In an interpretation-use argument, four assessment inferences are proposed (scoring, generalization, extrapolation and implications) each requiring various sources of evidence to assess the soundness of each proposition, followed by collection and examination of the evidence to determine to what extent it supports or weakens how assessment data are interpreted [Appendix B]. If a rating score is a fair, accurate, and a reliable reflection of an observation, then scoring inferences support the interpretation argument. If the items and other conditions sampled in the assessment represent the possible items and conditions that could be assessed in the test performance domain, then generalization inferences support the

interpretation argument. If the assessment provides adequate measures of “real world” performance, and is not overly influenced by extraneous factors, then extrapolation inferences support the interpretation argument. If the consequences of the assessment on learners, stakeholders, and society at large are favorable, then implications inferences support the interpretation argument [40]. Although all inferences merit attention, their relative importance depends on the context and specific interpretation-use argument [40]. Kane’s framework can highlight gaps in relation to certain inferences in the research literature, which helps appraise the strengths and weaknesses of assessment tools. The validity argument framework emphasizes the need to check our assumptions when we interpret scores, and allows for alternative interpretations and uses of assessment scores [53].

Our team’s understanding of how to utilize Kane’s framework is outlined below in eight general steps, based on work from Cook, Brydges and Hatala [Appendix D, 40, 44]. Our specific definitions and interpretation-use argument, are outlined in Chapter 4: Research Questions and Hypotheses.

- 1) **Define the construct being measured, and proposed interpretation** of the assessment data.
- 2) **Make explicit intended decisions** resulting from interpretation and use of assessment data.
- 3) **Define the interpretation-use argument** of assessment data which would support the decision(s), articulating hypotheses for key inferences (scoring, generalization, extrapolation, implications). Prioritize needed validity evidence according to key stated assumptions, and anticipated evidence sought [44]. Of particular interest are the weakest assumptions which

support or refute the hypothesized assumptions, as, observed by Clauser, “the validity argument is only as strong as the weakest link in the chain of inference” [53].

4) **Identify candidate instruments, or create/adapt a new instrument.** Instrument should align conceptually with the target construct. It is preferable to use existing instruments, which permits comparison with prior work, and includes evidence in the overall assessment base for that instrument, task or modality.

5) **Appraise existing evidence** for the key inferences (scoring, generalization, extrapolation and implications) from the literature published on the target assessment tool or program, and collect new evidence as needed. The direction and magnitude of evidence (favorable/unfavorable, and to what degree) should be noted, as well as what gaps remain, and for which contexts evidence is relevant. Researchers should avoid the tendency to focus on easily accessible validity evidence rather than the most important [44].

6) **Keep track of practical issues** in development, implementation and interpretation of scores, including cost and feasibility, which are important and understudied.

7) **Collect and synthesize validity evidence to formulate a validity argument, which is compared to the interpretation-use argument.** Most often these are not perfectly matched, and it is important to note where evidence may not be as favorable as expected, and where gaps exist, to guide future research. Iterative revision of the assessment tool instrument or proposed use may occur here, to address evidentiary gaps and respond to weakest assumptions.

8) **Make a judgment:** does the evidence support the intended use and to what degree? Are costs reasonable? Through this process, the validity argument is evaluated and the tool or proposed use may be provisionally accepted for use in a particular context.

3: REVIEW OF RELEVANT LITERATURE

3.1 Applications of Kane's Validity Framework in Medical Education

The concept of validity frameworks has been well articulated in the literature, but few studies have applied contemporary frameworks to assessment in medical education. A systematic review by Cook et al. of validity evidence for assessments of technology-enhanced simulation published before 2011 revealed 24% of studies did not reference a single validity framework, 3% referenced Messick's framework, and none mentioned Kane's framework [54].

Since the publication of this systematic review, five author groups have used Kane's framework to study the validity evidence for tools designed to measure various constructs in medical education (e.g., professionalism, surgical technical skills, medical school admissions knowledge) [55-59]. In these papers, the authors combed the relevant literature and weighed supportive evidence for scoring, generalization, extrapolation and implications inferences. This process revealed evidentiary gaps and provided a stimulus for future research directions.

Hawkins et al. applied the Kane approach to assessments based on the Mini-Clinical Evaluation Exercise (mini-CEX), an assessment where faculty observe trainees conducting a focused task in a clinical setting, and score on a global rating scale [55]. Scoring validity evidence was the weakest component of the validity argument, and limited by known problems with global rating scales including leniency error, high inter-item correlation, and inconsistent rater selection and training. Potential causes for this were postulated, and research aimed at clarifying further was

suggested. Generalization validity evidence drew from varying study designs and uncontrolled settings, but those in controlled settings yielded defensible dependability coefficients for eight to ten encounters. Extrapolation evidence was consistently supportive: mini-CEX performance differed by level of proficiency and/or training, and mini-CEX performance was done in the real-life practice setting. For the implications inference, authors noted associations for the extrapolation inference were theoretically plausible, but questioned whether educational interventions would increase low or domain-specific mini-CEX scores, and further study of current mini-CEX use in high stakes situations was valid, given that the mini-CEX was designed for formative use.

Boulet et al. used the Kane framework to think about and plan studies in simulation-based assessment, with emphasis on research [56]. The authors also used Messick's and an additional four-level training evaluation framework by Kirkpatrick, as complementary approaches to prioritize evidence. Boulet's review found that scoring validity evidence was often described incompletely, with lack of standardization and specialty-specific scoring models preventing evaluation of comparative utility. A paucity of generalization validity evidence was found, and research for how best to choose scenarios as well as how to measure, identify and minimize error was proposed. Extrapolation evidence was rare, and involved associating scores with performance in practice and readiness for advancement in training. Research needs included quantifying these relationships and identifying threats to validity. Implications evidence was solely described in terms of research gaps, including studies to substantiate the theoretical basis for what is measured to derive defensible standard setting methods and ideal structures for

educational assessments. Boulet concludes with consensus recommendations for future research in simulation-based assessment which relate directly to gaps identified in the validity argument.

Clauser et al. used Kane's framework to collect evidence to support the validity of assessments of professionalism in medical education [53]. Scoring validity evidence was difficult to find, as there was ambiguity in definition of professionalism, few instruments available, issues with rater familiarity bias, rater subjectivity, and considerable variability in scoring. A search for generalization evidence found that coefficient alpha was the most common measure of reliability, but this methodology was thought to be "misleading." Rarely, studies using generalizability (G-study) analyses provided results reflecting the contributions of rater and patient sampling to the variation in assessment scores. Extrapolation validity evidence was not readily available because external criteria for professionalism were difficult to identify, and the most common validity evidence was "expert" agreement that assessed domains were appropriate, which was considered weak in Kane's framework. Threats to validity evidence were discussed, which further challenged the extrapolation inference. The authors did not judge the implications inference as relevant in their analysis. Overall, the authors produced a validity argument suggesting that, regardless of the intended use of assessment scores, critical evidence was lacking to support high-stakes assessment of professionalism.

Hatala et al. used Kane's framework in a systematic review of validity evidence from Objective Structured Assessments of Technical Skills (OSATS), an assessment using direct observation of

multiple tasks rated by a task-specific checklist (CL) and behaviorally-oriented global rating scale (GRS) [57]. Scoring evidence was frequently collected and was usually favorable for the OSATS, likely because the construct of 'technical skill' is less ambiguously defined than constructs above. Rigorous CL development, rater training, and methods to reduce rater bias were areas for testing additional OSATS scoring inferences. No OSATS generalizability studies were found; however, inter-rater reliability of both CL and GRS were typically excellent, and thought to provide supportive generalization evidence. Extrapolation evidence was best described and well-supported, with consistent positive correlation of scores as expected with expertise, post-training status, and measures of surgical technical skills including real-world proficiency. The weakest evidence base was for implications, with no studies addressing the impact of the assessment although pass/fail standards were set in four studies. Hatala et al. concluded that the cumulative validity evidence seemed to reasonably support use of the OSATS as formative assessment, and suggested the need for researchers to explore feedback and relative utility of GRS and CL. For summative use and program evaluation use, OSATS use was thought questionable due to the dearth of significant evidence, particularly for the generalization and implications inferences.

Kreiter et al. reviewed the validity evidence for use of medical school admission testing using the standardized Medical College Admissions Test (MCAT), motivated in part, by untested assertions in medical education that MCAT testing may not effectively predict professional performance [58]. They studied assessments of general mental ability (GMA), a theoretically equivalent measure of cognitive ability from social sciences literature, closely associated with MCAT

performance. Rather than itemizing validity evidence, Kreiter focused on the studies most important to their validity argument: those investigating selection for employment. Meta-analyses drawing from a vast literature of psychological investigations demonstrated that GMA was a highly effective predictor of performance for virtually all occupations, particularly for intellectually challenging and highly complex professions, thus providing supportive evidence for the extrapolation inference. Kreiter also explored research implications, for example high GMA was associated with lower likelihood for counterproductive (in medical literature, “unprofessional”) work behaviors. Kreiter summarized that substantial evidence supported selecting for cognitive ability as a robust predictor of professional performance, thus refuting prior speculation.

Till et al. showed how assessment designers could use an argument-based validity framework practically, analyzing the validity evidence for a final-year medical student simulation exercise used to assess clinical ability as part of a portfolio of student’s capability as a practitioner, professional and scholar [59]. They collected evidence and developed a validity argument based on Kane’s inferences, though the validity argument was not explicitly stated. Sound evidence supported the scoring inference, including content alignment with behaviors defined through rigorous processes, attention to minimizing sources of error, authentic context, expert raters, and good reliability. Generalization evidence through multi-facet Rasch measurement and generalizability theory approaches demonstrated that the largest contributor to variance in ratings was systematic differences between students, and the exercise could reliably separate students into distinct levels of clinical ability. Decision studies (D-studies) modelled optimal

reliability by assessor and domain. Extrapolation validity evidence thus supported the interpretation, though the authors recommended further collection of correlational data. Implications evidence was reported as acceptable generalizability coefficients for high stakes decisions, but the authors did not consider potential unintended consequences of use. Till et al. concluded that while the proposed score interpretation was defensible for its purpose, if the exercise were to be used in a stand-alone fashion for a high-stakes decision, modifications based on their analysis would be necessary.

The above examples demonstrate different approaches to using Kane's validity framework to study distinct, unrelated constructs. Despite these differences, similarities were observed in synthesized research needs. Authors proposed researchers utilize more standardized instruments and study designs to permit comparative evaluation, to reduce reliance on global rating scales, to ensure consistent rater selection and training, to perform dedicated G-studies, to quantify associations between variables, to identify threats to validity, to derive methods for defensible standard setting, and to establish ideal structure for educational assessments. Authors emphasized that their critical appraisals yielded more than analytic content, suggesting the process they used could serve as a template for educators and future researchers to deepen understanding of interpretations made from assessment scores.

3.2 Applications of Kane's Framework to Echocardiography Competency Assessment

We used Kane's approach to critique the validity evidence for assessment of echocardiographic skills. The literature reveals a paucity of research in echocardiography skills assessment, an

echocardiography research agenda driven by non-cardiology subspecialties with different scopes of practice for this skill, variable definitions of the concept of “competence,” and no gold standard for assessing echocardiography skills development. Although two authors have sought validity evidence for echocardiography skills using Messick’s framework, there are no applications of Kane’s argument-based validity framework in the literature [60, 61].

To comprehensively review the validity evidence for echocardiography assessment skills, we thus expanded our literature search to encompass assessment of both echocardiography and general ultrasound competency. We reviewed the medical education literature according to principles of systematic review. This review, in full, is beyond the scope of this thesis, but the process and data yielded is appended in [Appendix E](#) and [Appendix F](#). Thirty-two studies of assessments of competence in echocardiography or general ultrasound were included [37, 38, 60-89], with scoring, generalization, extrapolation and implications evidence collated in a worksheet according to Kane’s framework, and judged in terms of the direction and magnitude of evidence (whether evidence was favorable or unfavorable, and to what degree). This process revealed evidentiary gaps in echocardiography and ultrasound competency assessment, and is the foundation for this thesis. Highlights of this analysis are summarized below.

Interpretation-use arguments in ultrasound assessment literature

An interpretation-use argument goes beyond a statement of purpose to articulate the assumptions to be confirmed or refuted in the ensuing study [44]. None of the studies specified an interpretation-use argument explicitly, as none used Kane’s framework. Most studies stated

a purpose, generally whether assessment performance adequately measured clinical competence [60-63, 65, 74, 77-87, 89]. A few studies stated their purpose was to seek validity evidence, though a framework was only utilized in two [60, 61, 69, 75, 77, 79, 80, 81-83, 85, 87]. A trend observed in authors' discussions was to extend interpretation and use of results for an assessment beyond the initial study context. These propositions run counter to Cook and Hatala's assertion that: "validity evidence applies only to the purpose, context, and learner group in which it was collected; existing evidence might guide our choice of assessment approach but does not support our future interpretations and use" [41]. This misalignment is a point of concern, as validity theorists note that "application of test scores beyond the scope of existing evidence constitutes... a misuse" [41], and could be viewed as "exploratory empiricism" [54].

Validity evidence for the scoring inference in ultrasound assessment literature

The scoring inference states that the grade or narrative comments produced based on an observation are fair and accurate, and adequately capture key aspects of performance. Scoring validity evidence was well described in the included studies and was generally favorable, though the strength of this judgment varied greatly depending on study design. The strongest evidence derived from studies in which assessment tools were designed using functional task alignment to professional national and international standards with expert input [36, 60, 61, 64, 65, 66, 68-72, 74, 75, 78, 85, 87, 89], and/or modifications of rigorously-developed scales [78, 79, 80, 81, 82, 83, 89]. Strong scoring validity evidence arose in study protocols optimizing inter-rater reliability through direct observation, automated measurements of performance [38, 63, 77, 83, 87], multiple raters, and attention to factors to reduce rater bias. Rater training was infrequent

(<16% of studies), but where done, was described comprehensively and yielded strong scoring validity evidence [60, 61, 75, 79, 81].

Insufficient or weak scoring validity evidence derived from studies which did not articulate rationale for content inclusion, did not pilot test materials, which utilized single and/or unblinded raters [66, 73, 76, 86], or where validity evidence for the proposed assessment tool was not considered prior to implementation [88]. None of the studies reported robust qualitative evidence to support scoring inferences, although a few studies reported responses to questionnaires where participants rated perception of confidence [64], utility of training [84], perceived barriers to training [72], and simulator realism [84], and where raters ranked fairness and provided informal feedback [61,72]. Future research focusing on scoring validity evidence would benefit from more prevalent, rigorous rater training and in-depth qualitative assessments capturing rater and participant feedback on the assessment and the overall educational experience.

Validity evidence for the generalization inference in ultrasound assessment literature

To satisfy the generalization inference, an assessment's sampled observations must be representative of all possible observations in a performance domain for a particular skill, and reproducible if a different sample were obtained. There has been little work done in the field of acquiring and reporting generalization validity evidence in simulation-based assessment [56]. In our review, we encountered challenges in synthesizing generalization evidence for echocardiography assessment. The first issue was that some assessment tools did not sample a

domain for a particular skill, but rather, represented the entire skill set an individual was trained to do, such as a detailed 88-item procedure specific transthoracic echocardiogram checklist [60], or the entire 37 required views on a perioperative transesophageal examination [64].

Conversely, there may have been a specialty-dependent factor, as other tools designed for anesthesia and critical care, did include samples of aspects of the abbreviated focused echocardiography protocol [37, 61, 67].

Generalization evidence from assessments for general ultrasound could not be readily compared with that obtained from echocardiography assessment due to the distinct differences in the domains sampled; however the evidence was more robust due to tools utilized in study design. Many general ultrasound assessments were derived from an observed structural assessment for ultrasound, which is tailored to use in an individual specialty by content experts who have consciously discussed and selected the skills sampled a priori to reflect the domain [78, 79, 81, 82, 83, 90].

A generalizability study, or G-study, can provide input to determine the required number of observations and reproducibility of individual observations, and thus provide a strong source of generalizability evidence. Two of the 32 studies reviewed [71, 81] performed G-study analyses with accompanying D-studies to model optimized feasibility and reliability. Future research to yield additional generalization validity evidence could consider generalizability studies in their protocols to reinforce sample size and statistical power decisions. Sampling strategy, particularly for echocardiography-specific tests, should be described and a rationale for whether it was

purposely reflective of an entire domain or intended to be representative. A qualitative component sampling the perspectives of the participants and raters and ensuring transparency in the interpretation would provide additional generalization validity evidence.

Validity evidence for the extrapolation inference in ultrasound assessment competence

literature

The extrapolation inference states that the scores generated in an assessment setting reflects meaningful performance in a real-life setting. This inference is vitally important in CBME, which seeks tools to link observed performance on assessments with professional clinical competence. Extrapolation validity evidence was the most frequently cited evidence in 28/32 studies (89%), which aligns with the literature on simulation-based assessment [54]. The extrapolation validity evidence presented was predominantly the positive correlation between expertise and performance [60-63, 65, 66, 74, 75, 77, 78, 79-84, 87], with some studies able to demonstrate this further by gradation in level of training [78, 79, 80, 81] across a range of assessment measures [60, 63, 66, 68, 74, 77, 79, 82, 83, 84, 87]. However, Cook and Hatala note expert novice comparisons can be confounded by unrelated factors and thus provides weak evidence which “adds little to the validity argument” [44, 91]. Similarly, in the subset of educational intervention efficacy studies where expert-novice comparisons could not be done (as there was a homogeneous population with one level of expertise), favorable, but weak, extrapolation validity evidence demonstrated a training effect. Specifically, testing before and after an intervention demonstrated improvement in a variety of circumstances and using a variety of measures. Again, this does not add significantly to extrapolation validity evidence.

Notably, there were occasional paradoxical results which failed to distinguish experts and novices [60,63,65,77,82,87,88], which evokes Cook's statement that "[results] are most interesting if they fail to discriminate groups that should be different, or find differences where none should exist" [44], but this also has the effect of reducing the strength of extrapolation validity evidence. Authors attempted to postulate reasons for these differences, for example proximity to ultrasound experience in the curriculum, or underpowered sample sizes not permitting intra-group comparisons.

Stronger extrapolation validity evidence was demonstrated by observed correlations between test scores and other measures having an expected relationship (criterion-referenced measures), for example a correlation between economy of hand motion metrics with scores for an objective measurement of focused assessment with sonography for trauma [83], or reduced angle correlated with cognitive skill for echocardiography [63]. What became evident in reviewing the literature, however, was the lack of an empiric echocardiography standard by which these assessment tools could be compared, with Millington et al. proposing a comparison of assessment scores with objective structured clinical examinations or mini-clinical examinations to examine relationships with other instruments [61].

In the ultrasound assessment literature, simulators were used for teaching and assessment, but there was variability in whether performance in the simulated environment was translatable to performance with real patients, and quantifying this relationship was not the focus of this review. Transfer between settings remains an interesting question and could be studied further

to obtain additional extrapolation validity evidence. The bulk of extrapolation evidence to date is easily-accessible, low-yield data which does not advance a validity argument. It would advance the field further to go beyond this, and look for more insightful correlations comparing plausible connection between skills and the real world for further research.

Validity evidence for the implications inference in ultrasound assessment literature

The implications inference states that assessment performance is used to make meaningful decisions with favorable consequences for learners, stakeholders and society at large. While this may be considered the most important type of validity evidence, it is reported infrequently [41] and was similarly observed sparingly in ultrasound competence assessment literature.

Six of 32 studies (19%) yielded implications validity evidence, most often with respect to standard setting to separate experts from novices and define cut points [75, 77, 79, 80, 81, 87]. One study defined a mastery learning target for longitudinal training [88]. As most instruments were novel, studies did not explore any of the following: comparison of actual vs. expected passes or failures; exploration of anticipated impact of testing on students, patients and raters; agreement of raters with final interpretations; or observed real-world test performance across learner groups. Participant and rater perspectives were reflected by questionnaire responses in eight studies of 32 (25%), but a detailed understanding of the accuracy, authenticity, fairness, and perceived impact of assessment testing was not provided. Millington suggested “future studies should compare scale ratings to... consequences of implementing point-of-care US assessment on educators and learners” [61]. Other authors suggested reviewing the role of simulation-based assessment to alleviate human resource reliance in training programs [77,87],

considering the use of the “cumulative sum of scores” in serial assessment to monitor formative progress [79], and thinking about how to establish the ideal length and structure of educational assessments [56]. These are important directions for developing future research and implications validity evidence.

In summary, using Kane’s framework to review the literature on assessment of echocardiography and ultrasound revealed the following gaps: few studies make use of validity frameworks. None of the articles utilized an interpretation-use argument, resulting in incomplete evidence, and occasionally a threat of extending the conclusions beyond what was observed. The scoring validity evidence appeared favorable overall, though there was variability depending on the individual study. More rigorous rater training and in-depth qualitative narratives to capture participant and rater perspectives are needed. Generalization evidence was heterogeneous, and would benefit from additional generalizability studies and explicit sampling strategies. The majority of extrapolation evidence was well-described, but insufficient and easily-accessible without adding significantly to the over-arching validity evidence for these tools. Better, more thoughtful measures that advance the interpretation-use argument are required. The primacy and paucity of implications evidence presents an opportunity for studies to better understand the role of simulation on human resource planning and formative cumulative testing.

3.3 Applications of Kane’s Framework in our context

Using Kane’s approach to critique validity evidence for assessment of echocardiographic skills revealed several areas where better quality evidence is needed. The scope of this thesis is to

build on the identified research needs, and to produce robust validity evidence for an ECAT in the context of the postgraduate cardiology training program at the University of Toronto.

To understand the needs in our context, we gathered anecdotal feedback from cardiology trainees in a town hall about their echocardiography training experiences. This needs assessment identified initial learning in the echocardiography lab as a curricular gap, with junior trainees reportedly feeling apprehensive to perform echocardiography studies on call. We engaged a range of local trainees, educational and echocardiography experts, in a modified Delphi process to determine learning objectives for core knowledge, technical and evaluation skills, and attitudes deemed appropriate for novice learners prior to echocardiography lab exposure. That process uncovered fundamental differences in opinion on whether trainees should master technical skills prior to patient exposure, although consensus was reached on key items with respect to knowledge, attitude, and evaluative skills [25].

We identified resources and formulated a plan to meet these needs. We obtained access to a high-fidelity echocardiography simulator on which trainees could practice performing an echocardiogram. We proposed that studying a modified assessment tool, used formerly on a standardized patient, to assess simulation-based basic echocardiographic skill performance would address our local needs, and contribute validity evidence to support or refute use in formative assessments of cardiology trainees.

4: RESEARCH QUESTION AND HYPOTHESES

4.1 Research Question

What is the nature of validity evidence collected when using the ECAT for formative assessment of cardiology trainees' simulation-based basic echocardiography performance, including their scanning skills, and identification of anatomic structures?

4.2 Interpretation-use Argument

We propose that a simulation-based ECAT can be used as a formative assessment, meaning that participants will use the feedback provided to formulate a learning plan for developing and refining their echocardiography skills, and that raters will use the ECAT tool to provide specific and actionable feedback. Beyond a formative use, we propose that participants' ECAT scores can be used as an additional data point in competency-based decisions, and will be positively related to ratings of their the global diagnostic quality of their simulated echocardiographic scans, to their summative echocardiography exam scores, and to their postgraduate training year.

4.3 Hypotheses

We expect the ECAT will have sufficient favorable validity evidence to support its use in formative assessment of cardiology trainees. See Table I for a rationale for prioritizing data collected for each inference.

Implications

Hypothesis 1) Participants will perceive that ECAT testing and feedback facilitate learning and improved echocardiography skills.

Hypothesis 2) Raters will perceive that ECAT testing and feedback facilitate learning and improved echocardiography skills.

Scoring

Hypothesis 3) Raters will use the ECAT consistently when scoring.

Hypothesis 4) Dimensions on the ECAT will be unique and demonstrate evidence of independence.

Extrapolation

Hypothesis 5) Participants' ECAT scores will be positively associated with observed performance on end-of-year echocardiography examination.

Hypothesis 6) Participants' ECAT scores will be positively related to the global impression of whether the simulated echocardiogram is of diagnostic quality.

Hypothesis 7) Participants' ECAT scores will discriminate trainees according to expected level of performance by their postgraduate training year.

Generalization

We did not investigate evidence related to generalization, as that was not a priority for the interpretation-use argument we articulated.

TABLE I – SUMMARY OF INFERENCE CATEGORIES, APPRAISAL OF EXISTING EVIDENCE, RATIONALE FOR PRIORITIZATION, CLAIMS AND HYPOTHESES, AND ASSOCIATED METHODS AND ANALYSES

	Definition	Appraisal of existing evidence	Why prioritized in this sequence?	Hypotheses and Claims	Methods and Analyses
Implications	Assessment performance has meaningful consequences for learners, teachers and systems.	Evidence is infrequently reported and solely for summative use, with standard-setting through method of contrasting groups.	In our interpretation-use argument, formative assessment and feedback have important consequences for learners and raters in a cardiology training program.	<u>Hypothesis 1</u> : Participants will perceive that ECAT testing and feedback facilitate learning and improved echocardiography skills. <u>Hypothesis 2</u> : Raters will perceive that ECAT testing and feedback facilitate learning and improved echocardiography skills.	For hypotheses 1 and 2, all participants and raters were interviewed. Using a simple content analysis framework we sorted, coded, and synthesized interview data into themes.
Scoring	The way an observation in an assessment is scored, including scoring rules, rubric and procedures, are done fairly and accurately, and capture key aspects of performance.	Strongest evidence reported from studies of tools with scoring rubrics established by rigorous methods and functional task alignment, stating how error was minimized, reliability maximized, and consistency emphasized. Rarely described rater training or supportive narratives.	Ours is the first study of a new tool, thus no previous evidence and the need to ensure scoring rules, rubric and procedures are fair and accurate.	<u>Hypothesis 3</u> : Raters will use the ECAT consistently when scoring. <u>Hypothesis 4</u> : Dimensions on the ECAT will be unique and demonstrate evidence of independence.	For hypothesis 3, we examined intra-class correlation coefficients. For hypothesis 4, we conducted factor analyses and calculated Cronbach alpha estimates.

TABLE I (CONTINUED) – SUMMARY OF INFERENCE CATEGORIES, APPRAISAL OF EXISTING EVIDENCE, RATIONALE FOR PRIORITIZATION, CLAIMS AND HYPOTHESES, AND ASSOCIATED METHODS AND ANALYSES

	Definition	Appraisal of existing evidence	Why prioritized in this sequence?	Hypotheses and Claims	Methods and Analyses
Extrapolation	Assessment performance provides adequate measures of how candidates will perform in clinical contexts.	Evidence cited frequently for other tools, most often noting scores discriminated levels of expertise (considered weak evidence), and correlate with other measures with plausible associations (e.g. economy of hand motion metrics with higher test scores).	A secondary focus; we had data we could use to relate ECAT score to potentially relevant metrics.	<p><u>Hypothesis 5</u>: Participants' ECAT scores will be positively associated with observed performance on end-of-year echocardiography examination.</p> <p><u>Hypothesis 6</u>: Participants' ECAT scores will be positively related to the global impression of whether the simulated echocardiogram is of diagnostic quality.</p> <p><u>Hypothesis 7</u>: Participants' ECAT scores will discriminate trainees according to expected level of performance by their postgraduate training year.</p>	<p>For hypothesis 5, we analyzed the correlation between ECAT scores and end-of-year cognitive assessment score.</p> <p>For hypothesis 6, we analyzed the correlation between ECAT scores and determination of clinically relevant diagnostic quality.</p> <p>For hypothesis 7 we analyzed associations between ECAT score and level of training.</p>
Generalization	Items sampled in the assessment protocol are representative of the theoretically possible items in a complete echocardiogram.	Insufficient evidence; Rare G-studies and D-studies which are tool- and interpretation-use argument specific.	Not a priority for our interpretation-use argument. For further details pls see discussion.	n/a	n/a

5: METHODS

5.1 Sample Size and Participants

The purpose of this study was to evaluate validity evidence for formative use of the tool, meaning sample size was based on convenience, and some previous literature. Given the proposed assessment tool was new, estimation of standard deviations or determination of a meaningful difference in test scores was not possible. We predicated decisions of reasonable sample size on the sample in which the tool was tested [60], and previous work using workplace-based assessment tools and observed structured technical examinations, which deemed sample sizes of 5-20 participants sufficient [93-95].

Fourteen current cardiology trainees at the University of Toronto, Ontario, Canada were recruited for this project. The sample reflects a spectrum of expertise, ranging from first year cardiology trainees who have never performed an echocardiography to senior echocardiography fellows pursuing advanced postgraduate training in echocardiography after completing six core months of echocardiography during cardiology residency training. The study was extracurricular, and no compensation, monetary or academic credit, was provided for participation.

5.2 Setting

Data were collected across fourteen individual sessions at the University of Toronto Centre for Excellence in Education and Clinical Practice between June and July 2016. Efforts were taken to

ensure the delivery of each session was identical (i.e. same location, equipment, staff, instructors, raters, interviewers, etc.).

5.3 Design

The study employed a mixed methods study design including a single iteration of assessment, participant and rater interviews after assessment testing, and a prospective observational component.

5.4 Materials

1) Data Collection Form

The data collection form included a unique participant study identification number, postgraduate year of training effective July 1, 2016, sex, free-text self-reported number of previous echocardiograms seen, done and interpreted, and number of general ultrasounds seen, done and interpreted ([Appendix G](#)).

2) Orientation

A standardized 15-minute orientation to ultrasound physics, basic echocardiography views, and the echocardiography simulator was developed (CAE Vimedix, CAE Healthcare Inc., Montreal, QC, Canada) using PowerPoint (Version 15.20, Microsoft, Redmond, WA, USA). We created a reference sheet based on teaching materials used with cardiology trainees at multiple institutions, indicating the sequence of images required in the complete transthoracic

echocardiography examination and expected anatomy visualized on each of the standard echocardiography windows ([Appendix H](#)).

3) Echocardiography Competence Assessment Tool (ECAT)

There is currently no “gold standard” for assessment of echocardiography skills in training. The ECAT arises from a tool proposed for clinical echocardiography evaluation over a spectrum of expertise in standardized patients, modified per the objectives of this study [60, [Appendix I](#)]. The benefit of using existing instruments to assess most constructs for learner assessments as summarized by Cook and Hatala, is this allows comparison with prior work, “permits others to compare their work with ours, and includes our evidence in the overall evidence base for that tool, task, or assessment modality,” thus addressing evidentiary gaps in the literature [44].

Nielsen’s original tool [60] used a five-point Likert scale to qualitatively assess technical proficiency from very poor to very good for 88 dimensions in nine standard echocardiography views; we re-aligned rating to descriptive behavioral anchors for competencies expected at different levels of residency training, based on the ACGME internal medicine subspecialty and CoCATS4 Task Force 3 Milestone rubrics: not done (milestone 0-12 months or novice), expected performance by the end of PGY 4 (milestone 12 months or early learner), expected performance of PGY 5-6 (milestone 24-36 months or advancing learner) or expected performance at end of PGY 6 (milestone 36 months or ready for unsupervised practice/aspirational) [10, 24]. Points were allocated for each milestone, with “not done” assigned 0 points, “expected performance at the end of PGY 4” assigned 1 point, “expected performance of PGY 5/6” assigned 2 points, and

“expected performance at the end of PGY 6” assigned the highest number of points, 3 points.

Our proposed tool was reviewed for clarity, relevance and accuracy by an expert panel of National Board of Echocardiography-trained echocardiographers, and we made revisions based on their feedback.

Nielsen’s tool [60] required individuals to achieve 88 echocardiography images in nine standard echocardiography views recommended by the Danish Cardiac Society standard echocardiography examination; we reduced this to 24 images for the 12 standard transthoracic echocardiographic views per American Society of Echocardiography standards [29]. We decreased the number of required performance dimensions assessed from 88 to 24, based on the judgment that several of the images required in the original tool were highly technical. For instance, changing the sweep speed, gain and frame rate for continuous-wave doppler across the pulmonary artery in parasternal short axis, were a) not thought to be relevant to assessment of basic echocardiography technical and evaluative skills in trainees and b) were sophisticated measurements the high-fidelity simulator was either incapable of, or performed incompletely. The two performance dimensions assessed for each of the 12 standard echocardiography views were relevant echocardiography skills expected of all trainees: can you obtain an image, and can you identify anatomic structures [30]. Nielsen had used the term “anatomical representation” as a dimension but not clearly articulated what this would entail [60]; therefore in our rubric we specified “identification of anatomic structures.” Through this process, scoring items were reduced from 88 to 24 and the maximum total score was reduced from 440 to 72. In this way, we addressed a limitation Nielsen identified with their tool, in that a single rater with familiarity

with the long list of items had previously scored all 45 assessments in their study. Our simplification of the tool increased feasibility for raters and for overall implementation.

Nielsen's tool included a GRS to rate the overall quality of the scan on a five-point Likert scale, ranging very good to very poor, however the corresponding G- study demonstrated one third of GRS score variance was due to factors other than participants' expertise, calling into question the soundness of interpretations based on GRS scores [71]. As an alternative, we replaced the GRS with a global, binary determination of whether raters deemed the scan of diagnostic quality (yes/no). This score was intended to be clinically meaningful, given scans performed by trainees on echocardiography rotations and on-call after-hours become part of the official patient record, and the images are independently, often asynchronously, utilized by echocardiographers when generating clinical reports. As supporting evidenced, the O-SCORE included a similar binary score of procedural competency which was highly significant, and identified as a potentially useful endpoint for cumulative score time-trend analyses to monitor trainee progress [95].

We added a feedback component to the tool. Feedback is critical in formative learning as Norcini states, "for a test to provide effective formative assessment for the learner, it should provide specific and actionable feedback, be integrated into the learning experience, and be timely and ongoing" [96]. Most studies of tools have focused on objective outcomes, with only limited and highly variable discussion of feedback quality [56, 97, 98]. The OSCORE researchers responded to this by including two open ended questions for feedback in their rubric, which yielded important insights including that the assessment helped participants identify areas to

improve to become competent to perform a procedure independently, and raters found the assessment was practical and useful [95]. We consequently added two open-ended questions on the ECAT, asking raters to document positive aspects of the scan, and to provide suggestions for improvement, which would theoretically collect information to identify individual areas for improvement and identify process weaknesses.

Finally, we modified the setting and modality for the assessment. Nielsen's tool was used to rate performance on a single standardized patient who underwent 45 sequential echocardiography scans using the same clinical echocardiography machine, with ad-hoc, non-standardized assistance offered by a technical operator. To maintain standardization while improving feasibility, we altered the modality of assessment to a technology-enabled simulator (CAE Vimedix, Montreal, Quebec) programmed to normal, healthy mode. Technical assistance was offered as needed in practice prior to assessment but not during the assessment. Two other divergences from the tool protocol included that we provided a reference flowsheet of the standard echocardiography examination to practice prior to ECAT testing, and then we prompted for specific views in the same order during the assessment.

4) Echocardiography Simulator Video Clips

We exported de-identified three-second digital video clips of each participant's attempted standard echocardiography views directly off the echocardiography simulator immediately after testing, coded by the individual's study identification number.

5) Anatomy identification Video Clips

We filmed participant performance using a mounted wearable video camera (GoPro Hero 4, GoPro Ltd., San Mateo, CA, USA) from a single perspective, focusing on a frozen image on the simulator screen, with the only identifier being the individual's study identification number visible in the frame. A stylus was employed to point at a specific location on the screen and identify the anatomic structure. Auditory input was recorded simultaneously using a microphone. Using Apple iMovie (Version 10.1.3, Apple, Cupertino, CA, USA), we exported raw footage immediately after testing, and the de-identified digital files were coded with the individual's study identification number.

6) Participant and Rater Video Portal

A web-based password-protected portal was created for participants to access all digital video clips and for raters to access digital video clips and score a web-based version of the ECAT ([Appendix J](#)). After each individual testing session, the 12 de-identified echocardiography simulator video clips and 12 de-identified anatomy identification video clips were uploaded to a folder labelled with a participant study identification number. Participants could access their own personal folders as many times as they required. Raters could access participant folders for all participants who had undergone assessment testing.

7) Interview Guides

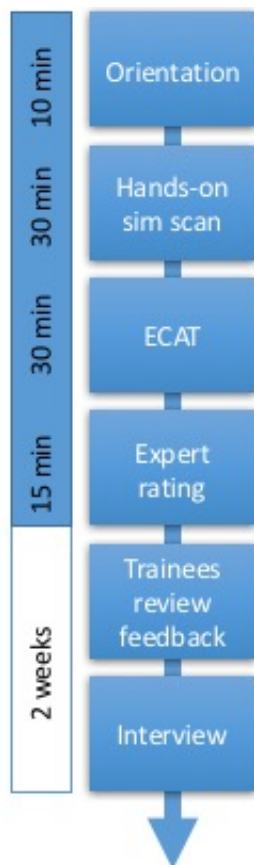
Interview guides were designed for raters and participants at completion of their respective data collection periods ([Appendix K](#)). Interview questions asked of participants and raters probed

their perceptions of the ECAT and feedback given, and to what extent feedback had an impact on practice behavior and echocardiography skills.

5.5 Procedure

Informed consent was obtained in compliance with Toronto Academic Health System Network, University of Toronto Office of Research Ethics, and University of Illinois at Chicago Institutional Review Board policy. Participants provided basic demographic information including level of cardiology training, number of previous echocardiograms seen, done and interpreted, and number of previous general ultrasounds seen, done and interpreted.

The study protocol is depicted in Figure 1. All participants received a standardized, one-on-one, interactive 15-minute orientation, which covered content areas of basic ultrasound physics, the twelve standard echocardiographic views comprising a complete echocardiographic examination, and the echocardiography simulator. We emphasized in the orientation and again, prior to the ECAT, that the intent of this tool was for formative purposes and to emphasize accuracy and quality over speed. This statement was also posted in ECAT testing instructions in the orientation room and above the simulator.

FIGURE 1: ECAT Study Protocol

Participants were subsequently provided with a reference sheet of the ASE standard echocardiography examination, adapted from a document provided to trainees at multiple institutions, which indicating the flow of the standard echocardiography exam, depicted ideal images, and indicated expected anatomy seen therein ([Appendix H](#)). Each participant had thirty minutes of hands-on time to practice acquiring standard views on the simulator programmed to normal, easy patient mode. The simulator was set up with the main screen demonstrating two-dimensional grey-scale images generated from the simulator echocardiography probe, with the

adjacent screen demonstrating three-dimensional human anatomy with real-time, dynamic plane adjustment and labelled structures. A technical assistant not involved with rating (G.K.), was present to assist participants with functions of the simulator and demonstrate how to toggle between the three-dimensional anatomic assisted mode to purely two-dimensional echocardiography mode if requested.

ECAT assessments commenced after thirty minutes of hands-on practice. The simulator was switched to two-dimensional grey-scale echocardiography mode and the reference sheet was removed. The technical assistant read aloud ECAT testing instructions posted beside the simulator. The participant was positioned on the right side of the simulator mannequin, as is standard in clinical echocardiography. One real-time rater with an ECAT scoring sheet was positioned unobtrusively behind the participant, to the left of the mannequin, with unimpeded visualization of the screen. The technical assistant turned on the mounted video camera, and prompted individuals sequentially for each of the twelve standard echocardiography views as per the score sheet, stating, "please demonstrate the parasternal long axis view." For each of the twelve standard echocardiography views, when individuals felt they had achieved the optimal view, they announced, "acquire," and the technical assistant pressed a button, acquiring a digital video loop of three consecutive cardiac cycles recorded from the simulator. The assistant then paused the participant's image and prompted participants to verbally identify anatomic structures visualized on the screen by pointing at them with the tip of the stylus. It was determined during preliminary testing that the suprasternal echocardiography window was

unattainable using conventional echocardiography maneuvers on the simulator, thus this window was excluded. Thus, we analyzed data from eleven standard echocardiography views.

Immediately following the session, de-identified digital video data from the simulator data and wearable video camera was downloaded, coded using unique numbers provided to participants on study entry, uploaded onto the web portal, and sorted into unique folders labelled by identification number. Backup copies of the coding sheet and a hard-copy external hard drive of de-identified digital data were stored in a locked office.

Rating was done in real-time and off-line by four National Board of Echocardiography board-certified expert raters (G.N., J.E., E.K., S.R.). All raters were blinded to participant identity for scoring, as participants could only be identified in real-time by their unique identification number on the echocardiography screen, and the digital data reviewed off-line was identified only by participant identification number. Furthermore, real-time rating was done in-person by one rater (E.K.) who was never involved in residency training and so was unfamiliar with the visual identity of participants. Real-time scoring and written feedback provision was done manually during the ECAT test, without verbal feedback or in-test interaction between rater and participant. Off-line rating was done by three raters (G.N., J.E., S.R.) reviewing de-identified video clips identified by study identification number on a password-protected web portal.

Anonymous, collated written feedback from real-time and off-line rating was compiled in a feedback folder on each participant's password-protected web portal. Participants could

privately log in at any time to view their individual simulator and anatomy identification video clips alongside written feedback. Participants never received the ECAT numerical score generated for their performance.

Two weeks after assessment, participants underwent a semi-structured telephone interview to determine if ECAT testing and feedback prompted changes in behavior or skills development, if they encountered barriers, and if they had suggestions to enhance feedback. To minimize any power differences, interviews were facilitated by an investigator not involved in echocardiography education curricular teaching. To maximize reflexivity, the interviewer used open-ended questions as initial prompts, and followed the interview script closely with relevant prompts and probes, trying at all times to facilitate participant interactions and avoid injecting their own biases into discussion [99]. The principal investigator interviewed raters at the conclusion of the study to examine their perspectives on to what extent ECAT testing assisted them in providing formative feedback, if they encountered barriers, and if they had suggestions to enhance feedback. We recorded audio from interviews, which was transcribed by an independent third party and analyzed in a de-identified fashion.

5.6 Rater Training

Echocardiography experts were selected as raters, as it is believed experts understand what is required in a competent clinical performance and can judge the quality and appropriateness of trainees' practice [95]. Raters were four experienced National Board of Echocardiography-certified individuals. Raters were both sonographers and cardiologists, representative of

personnel involved in academic echocardiography skills education in cardiology residency training. The assessment tool was distributed in advance, and usability was discussed. Two individuals at either end of the expertise spectrum were recorded performing three views, and simulated echocardiography video clips and anatomy identification clips were reviewed independently by all raters. At a one hour, real-time meeting, raters reviewed how they would rate each video clip using the ECAT, and discussed areas of agreement and disagreement to develop a shared mental model. Raters provided feedback for formative purposes, emphasizing specific aspects of performance, suggestions for improvement, and use of feedback to develop an action plan. On the basis of discussion in rater training, behavioral performance anchors were refined until coming to consensus as the final ECAT in our study ([Appendix I](#)).

5.7 Analysis

Analyses were aligned to hypotheses for the validity argument inferences ([Table 1](#)). A statistics consultant compiled all participant demographics and performance data in a database using Statistical Analysis Software version 9.2 (SAS Institute, Inc., Cary, NC). We aggregated the interview data, and did not link the data to any individual.

To obtain validity evidence relating to the implications inference, we explored participants' and raters' perceptions of the benefit of formative feedback generated during this assessment process. We used NVivo software (QSR International, Melbourne, AU) and a simple conventional content analysis framework to sort, code and refine qualitative interview data into emergent themes [100]. Initially, two investigators (S.R. and R.B.) independently coded two randomly

selected transcripts. During coding we read the transcripts multiple times, looking for patterns and/or unique insights in the data. Coding was inductive as we allowed codes to emerge as we read the transcripts, and deductive, as we used principles from what is known in the domain of feedback and self-regulated learning, to inform our interpretation. We met to review each member's code and went through this process four times until we could agree on a final grouping of codes judged to sufficiently represent the data. We maintained an audit trail including research meeting notes, individual's preliminary coding, the evolving grouping of codes into themes, and our decision process for choosing representative quotes. We treated the data comprehensively, meaning we aimed to account for all the codes we generated in the final thematic structure.

To analyze validity evidence related to the scoring inference, we computed test-item statistics including mean and standard deviation for each of the two dimensions assessed per echocardiography view. Internal consistency of the tool was assessed by coefficient alpha. Inter-rater reliability was assessed using the intra-class correlation coefficient (ICC) computed as an item-specific result and overall, with multiple models of rater subsets to determine interaction between severity of rater and overall score. An exploratory factor analysis was used to examine the relationships between the different echocardiography views and performance dimensions.

To analyze validity evidence relating to extrapolation inferences, univariate linear regression was used to assess relationships between participants' ECAT scores (the dependent variable) and

their self-reported sex, level of training, and prior experience with ultrasound and echocardiography. A generalized estimating equation model was used to find the association between score and binary determination of diagnostic quality. The relationship between average ECAT score and end of year echocardiography exam was assessed by correlation. For all assessments, an alpha level of $p < 0.05$ was considered statistically significant.

5.8 Reflexivity

As the principal researcher, I consciously tried to reflect upon the acquired data which I collected and analyzed. To enhance my reflexivity [100], and avoid my own assumptions and behavior impacting the inquiry process, an interview script was carefully designed with prompts to facilitate participant interactions. I worked closely with my co-investigators to develop this script, and we created probes to facilitate discussion while avoiding injecting personal biases into discussion. I tried to adhere to the script, prompts, and probes. I analyzed and interpreted data which was de-identified from the source interview material, in an unbiased fashion. Throughout the analyses, I tried to be aware of my own personal assumptions which might affect results.

I was ideally suited for this role, as I did not have a formal role in trainee echocardiography teaching. I was external to the hospitals in which the trainees rotated for their echocardiography training blocks, and did not encounter the residents on rotation. I did not do my cardiology training at the university where the study was undertaken, and had no preconceived notions of strengths or weaknesses of the existing system, save for what was relayed in the interviews.

6: RESULTS

6.1 Demographics

Our cohort included five first-year cardiology trainees, five second-year cardiology trainees, and four third-year cardiology trainees or postgraduate fellows (Table II).

TABLE II – PARTICIPANT DEMOGRAPHICS

Variable	Number
Participants, by level of training	n = 14
Cardiology year 1	n = 5
Cardiology year 2	n = 5
Cardiology year >=3	n = 4
Sex (female)	6
Variable	Mean (Standard Deviation)
Number of echocardiograms seen	306 (443)
Number of echocardiograms performed	100 (138)
Number of echocardiograms interpreted	255 (447)
Number of ultrasounds seen	30 (24)
Number of ultrasounds performed	14 (15)
Number of ultrasounds interpreted	8 (15)

6.2 Evidence for Implications

We explored perceptions of trainees and raters to see if ECAT testing and feedback facilitated learning and improved echocardiography skills. Review of interview transcripts from fourteen trainees revealed three major themes relating to implications of feedback for formative assessment. We came to consensus on three themes as representing the data, which are

summarized in **Table III**: i) feedback stimulated change, ii) how feedback was delivered impacted participants' perceived learning, and iii) assessment credibility influenced participants' receptivity to feedback. These themes reflect participants' beliefs regarding the role of feedback, in this study, and in their training to date. We describe the themes in more detail below, with representative quotes embedded in the summary syntheses.

TABLE III – LIST OF INITIAL CODES AND FINAL THEMES

Initial Codes	Themes
<ul style="list-style-type: none"> - feedback stimulated action - feedback was incorporated into self-assessment - issues with feedback recall and action plan 	Feedback stimulated change
<ul style="list-style-type: none"> - timing of feedback delivery: concurrent, terminal, through productive failure and incremental improvement - mode of feedback delivery - content of feedback and conceptual knowledge provided a basis for skills development - learners' theories regarding feedback 	How feedback was delivered impacted participants' perceived learning
<ul style="list-style-type: none"> - rater credibility - assessment process credibility - simulator credibility for task of basic echocardiography skills 	Assessment credibility influenced participants' receptivity to feedback

*codes not listed in any particular order

Theme 1: Feedback stimulated change

Almost all participants indicated that feedback they received through the ECAT process stimulated some type of action. Several participants could recall specific, technical feedback provided, and mindful action when performing echocardiograms going forward. For example, one participant commented:

“there’s a lot of things that I think I need to, in the future whenever I do an echo[cardiogram], I should focus on it, especially how to avoid the foreshortening the apex. At least I’ll keep it in mind so in the future, with my next rotation I will try to focus on these things even I might, for example, ask the sonographer to look at it before I save it, just to make sure I’m doing the right thing.” (Participant 13)

Many used feedback as a motivation for self-regulated learning and for goal setting. For example, a typical response was: “I think, number one, it made me go back to more first principles and look up things in textbooks to try and identify more of the anatomy and structures in preparation to perform the procedure.” Having access to a record of feedback, both visual and written records, permitted the opportunity for participants to review their own performance, relevant feedback, and incorporate into action plans moving forward. A participant felt,

“I think that it's pointed out some things I needed to look up. I sort of went through the images myself and just re-familiarized myself with different structures and things like that, in order to be better at reading it going forward.” (Participant 11)

Participants’ comments provided a glimpse into self-awareness and how feedback was incorporated into self-assessment. Sometimes they commented on their own performance according to their perception of performance of someone at their level of training. One participant stated, “on the day of, I was pretty proud that I, as a C1, I could get a lot of the images. But then I realized they definitely could be optimized further.” Another participant felt,

“that’s the level I wanted to be at. I knew I wouldn’t have an issue with obtaining the windows. It was more just getting more specific feedback on optimization. And that was in line with how I thought the feedback would be.” Another participant perceived feedback assisted in prioritizing learning needs, stating “it identified maybe some gaps in knowledge that would otherwise would not really be identified.” One participant attributed feedback to helping him/her become “more aware... I think I kind of have made it my own personal goal to do that.”

A minority of individuals had paradoxical or negative impressions of feedback stimulating change. One participant stated the feedback received was useful, yet could not recall specifics, “I’m trying to remember what exactly the detailed feedback was, but the constructive feedback I remember receiving on the day of was thinking, you know, these are things that I should work on.” This quote challenges the usefulness of the feedback, as such lack of recall would prevent the feedback from having a meaningful impact or leading to a specific action plan. Another participant was frustrated by the feedback received, as they had not yet undergone any echocardiography introduction, and were discouraged about technical aspects of the feedback and what it meant for them. This participant stated the utility of feedback for formative assessment might be downstream after introduction to echocardiography training.

Learning stage seemed to play a role in action plans, as some self-described senior learners hypothesized that the feedback received would have led to specific action plans if they were in an earlier part of the growth curve. Others earlier in the training program stated they had definitely used the feedback received on call, were trying to do informal scans when possible,

and felt they had a “head start” going into their echocardiography rotation. Feedback from one early learner however contrasted with the proactive nature of others, and highlighted the different motivations even within a training cohort. This participant hypothesized their echocardiography skills would decrease between echocardiography rotations through “disinclination to practice”, and that feedback received would be forgotten in this absence of active skills practice.

Relating our evidence back to the implications inference, participants generally did feel that ECAT testing and feedback they received stimulated specific, mindful action. Participants were able to benchmark their level of performance. Participants reported internalizing the feedback, and using it as motivation for self-regulated learning and goal setting. These findings suggest that ECAT testing and feedback facilitated learning and improved echocardiography skills. No unintended or harmful impacts were seen with respect to feedback stimulating change.

Theme 2: How feedback was delivered impacted participants’ perceived learning

Data coded under this theme related to participant comments on when, how and what type of feedback was delivered, and the perceived impact of these aspects on their learning.

Several participants commented on the learning implications of the different timing of when and how they received feedback: i) concurrently during the pre-ECAT introduction to scanning by a written cue flowsheet and haptic feedback, ii) terminally from ECAT raters after performance, and iii) concurrently from repeated productive failure on the simulator. According to

participants, feedback was felt to be “timely” and “fresh in [my] mind”, which helped with recall of self-performance and the feedback received. The slight delay in terminal feedback due to study design was deemed to be desirable by some, providing an opportunity to go back and review self-performance without being emotionally affected in the moment, as reflected by this comment, “I have to say a lot of the times when something is said to you without being able to go back and look at it and make sure you fully understand where you went wrong or what you labelled wrong, is not really useful.”

However, many individuals desired more concurrent feedback: “It would be really nice to have live feedback as you’re doing it because you can really ingrain, especially when it comes to optimization, really ingrain the types of motions that you need to have to try to optimize things.” Another participant remarked on the practical challenges: “the pros of immediate feedback would be the immediacy of it...I think the cons are the practical implementation of directed feedback i.e., there's 24 hours in a day, docs already don't have time.”

One participant used an analogy of sports training to integrate both concurrent and terminal feedback over time: “echo[cardiography] is like a skill like a sport. Like anything else, and I think to that extent in the same way as coaching for a sport... it would work exactly that way. That kind of feedback right after you've done something helps you to form a more conscious approach to the next time you do that same thing again.” Another individual suggested the learning benefits could accrue through feedback over time, “if there was an actual structured curriculum where there’s going to be a certain degree of progression and learning objectives,

progression and progressively higher level skill sets. I think people like seeing that they're getting better and they like knowing that there's a road map.”

A minority of participants, however, did raise the issue that they would have preferred immediate feedback on technical image acquisition skills, skills which they felt were more amenable to concurrent feedback. In addition, one participant was discouraged by the written feedback:

“To be honest this feedback was a little bit hard to interpret. There was a lot of technical information in it that I don’t necessarily know how to with improving my skill set versus if someone told directly, “this is what you’re doing wrong, move it this way, move it that way”. I’m actually finding it difficult as I have this stack (of written feedback) in front of me, I’m finding it difficult to digest the specifics of how to improve my technique In the moment. What I should change and what I did wrong 'cause it's real time, there's two way communication, there's correction of mistakes right away instantaneously. That would have been preferred.” (Participant 3)

Participants noted the multiple modes by which feedback was delivered, and perceived this to be beneficial. Images and feedback could be directly compared. The majority of comments were similar to this participant, who stated:

“I think what’s nice is that we’re getting the videos and we’re getting the written format. So it’s not like I just get the written feedback and then I’m like, wait a minute, what did my parasternal long (axis view) look liked again. So that’s a bit of a difference. I thought

that was really helpful. And I felt like it was really detailed. So while I think real-time feedback would be helpful, I think in a way you do kind of overcome that because you provide us the videos as well. So I'm overall happy." (Participant 12)

Participants also mentioned haptic feedback as the basis for conceptual knowledge. The simulator was thought to be very helpful in providing a mental model which was the basis for skills development:

"Even though it's an image on a screen, it's a bit more of a 3D of what you're looking at and why you get certain cuts of the heart. So when we first start echo[cardiography], a lot of it is just memorization rather than really where the probe is and how it's cutting through the heart whereas this program gets you to understand the three dimensional aspect of echo[cardiography] and why your 2D image looks the way it does. By in your mind knowing where you are, and how you're cutting the heart, and I think that makes a huge difference. Whereas you don't get that on a normal human being, obviously because you're not seeing much inside the body, you're just seeing the image. This is a bit different so at least you know why you're doing the things you're doing and why certain structures come first in your picture, whereas you don't really get that sort of feedback. I know when I first started Echo[cardiography], a lot of people tried to show me on a heart - they had an actual constructed heart made out of plaster to look at the images and kind of tell you. I don't think it's as good obviously, when you have just have this artificial heart there and they're like, "We're cutting it like this", and "We're cutting it like that". This is, in real time you can see where your probe was and where it was cutting and you

could see on the other screen what the heart is looking like in the body and how this image is being produced, I think it's a bit better.” (Participant 11)

Related to the benefits of the simulator, another participant commented:

“I really liked the fact that the simulator, you know, you have your echo[cardiography] images. But then you also have, side by side to that, initially when I was practicing, that other computer screen where it shows you the heart and where the probe is going. I find that’s very helpful. I think a lot of times with residents, the number one problem with ultrasound is orientating yourself and saying to yourself, ok, if my beam is shooting this way why am I seeing the structures in this orientation. At least that’s what I felt and a number of my friends have felt. So I felt the fact that we have both screens initially, just to learn from them and say to ourselves, okay I’m shooting the probe this way, these are the structures I’m seeing, this is why the cut: I found that really helpful.” (Participant 12)

In summary, participants favored the multiple modes of feedback received, but had mixed sentiments about the timing of the feedback, which could impact whether ECAT feedback had consequences on their learning. Individual preference ranged from concurrent feedback to a slight delay in terminal feedback or feedback through productive failure over time. A mismatch between a participant’s preferred timing of feedback and that of the study protocol could attenuate the impact on learning, as seen in the unintended impact of feedback apparently discouraging one participant. Participants felt additional haptic feedback from the simulator was a particular benefit to them to develop a mental model for skills development. The balance of

these findings suggest that the type and format of feedback had variable effects on whether ECAT assessment and feedback facilitated individual action towards developing echocardiography skills.

Theme 3: Assessment credibility influenced participants' reception of feedback

Participants perceived ECAT feedback was legitimate, based on their perceptions of raters' credibility, of the assessment process as fair, and of the simulator's realistic representation of the echocardiography tasks.

Although participants were rated by sonographers and echocardiographers, the participant consent form simply stated that an educator would provide constructive feedback. Participants received anonymized, collated expert feedback. When participants were notified in the interview about the different professional identities of raters, some valued this, saying, "I think it is good to have different perspectives because of the different skillsets," and:

"I think it's important to get both viewpoints because, I think, both staff [cardiologists] and sonographers might approach it in a sense, from a different way. Like, in the end, you do the same thing because everybody wants to acquire good images and the appropriate structures in each view. But I think it's nice to get feedback from sonographers and staff just because they do have different expertise and different advice to give depending on that expertise... I think it was more beneficial to me, in the end." (Participant 12)

Others felt as long as experts were involved, it did not affect the feedback they received or the way they interpreted it:

“Honestly, I don't know if it mattered who was giving the feedback as much as it was just getting the feedback itself. So the fact that the people were echocardiographers - I assumed that whoever was doing the feedback knew exactly, and more than I did obviously, so I don't know if I even paid attention to who exactly was giving me the feedback.” (Participant 11)

Participants described the assessment process using words like: fair, organized, well-structured, reasonable, realistic, and comfortable. A few participants articulated self-awareness around formative testing, for example:

“it's kind of interesting, like it felt like a test, so like any test you kind of think, if I was better prepared for this I probably could have done better. But if this is a cross-sectional assessment of how I'm doing, you know, I felt I'm doing pretty good... the natural tendency is to try to do that best that you can on every test, but I think the more important thing was... you kind of just let it free flow and said, just do your best at getting these views and then calling what you see.” (Participant 4)

Just one participant stated they did not find assessment process beneficial, stating, “I was just struggling. It would have been a lot easier if someone had said, “just point this way and do that and you'll find it and that's what you look for.” (Participant 3)

Participants described the simulator as both beneficial and detrimental to different aspects of echocardiography learning. The majority had never worked with an echocardiography simulator and commented on its “coolness” factor, realism and role in assisting their visual-spatial conceptualization. Participants felt the standardization of simulation would improve early learning: “the patient is going to be variable whereas this mannequin is more standardized and...it's just easier and more standardized in terms of getting down the basics.” However, some individuals, particularly those with more experience, felt:

“it was bit unnatural because the machine...the mannequin itself, in real patients there's a huge grey zone in between having a perfect image and seeing nothing, whereas in the mannequin it was more or less all or nothing. Either you got a perfect image, or you got nothing at all depending on how you move the probe. So I found that was a bit challenging and counterintuitive because normally, at least the way I scan is if you don't see anything, then you slide a little bit to the left or slide a little bit to the right, I slide up and I slide down and then you're able to tell whether it's getting better or worse. That day it was like, one moment you could see it and then one moment you completely couldn't.” (Participant 14)

Others did not feel the simulator challenged them enough compared to the clinical environment:

“ I guess the only thing that the simulator can't provide is all those patient factors, right? When you're not getting a clear window how to sharpen your image or how to adjust yourself to get it more in a better view, I guess, that's harder.” (Participant 1)

Most participants believed the ECAT process and expert feedback process were credible. One participant felt the assessment process was a struggle, and this may have been an unintended barrier to his/her learning. The credibility of the simulator was called into question by more experienced participants, but this did not seem to be an issue for novice participants, who felt practicing on the simulator conferred benefits for their learning.

Rater Perceptions of ECAT Testing and Feedback on Learning and Skills

Interviews with three raters yielded distinct perspectives on ECAT testing and feedback. Rather than selecting quotes across all raters, we explored each rater's perspective in depth.

Rater 1 felt that the ECAT testing process facilitated learning and improved echocardiography skills, often comparing the process with standard teaching. The rater attributed a large role to facilitation of immediate feedback:

"I think that one of the aspects that a simulator can do, that you can't do with a patient the same way, is... immediate feedback. So it often would be an uncomfortable situation for a real-life patient if [the trainee doing the echocardiogram] were getting direct feedback continually throughout the scan... which really is what is required for a novice beginner in their first exposures. It can be unnerving for a patient to go through that process.. [here] they can do it in a way where the resident is not uncomfortable because there's not a patient in front of them... self-directed learning can occur as well with a simulator which you can't do with real-life patients in the same way."

Rater 1 stated feedback was enabled through standardization of the simulator, as opposed to having clinical variability affecting feedback given in a real-life assessment, saying “in the course of a rotation, you work throughout the rotation with them, so you have some biases that are formed, and will influence the process... in one sense that is actually helpful, because it’s more – you're blinded, like you know, affected emotionally by watching them do it, or having that kind of interaction; it’s very concrete what you see.” Rater 1 described using the ECAT tool as a feedback template, identifying trainee strength and weaknesses in the dimensions assessed, and found that it “allowed me to formulate their performance... with a very practical approach of not only where they are at now, but what they could do moving forward to improve.” Rater 1 used clinical experience to translate the observation into an assessment, stating, “I used it as a stepping stone, so being able to actually take it from a more concrete numerical value, having it in that objective sense, you could then interpret it then within your own clinical knowledge, to translate it to the narrative.” Rater 1 identified a challenge to providing feedback as only being able to see the end-product of the assessment, without observing the process of optimisation; however they offered that this permitted a “less emotional” interpretation not subject to biases formed that “influence the process” when assessing an individual a rater has worked with through a rotation.

Rater 2 focused on use of the ECAT to provide structured feedback. At various times in the interview, the rater described the tool using terms such as “very, very helpful”, “easy to use”, “easy to read,” and “objective” like we have a list with what we should expect them to do”.

Rater 2 expanded on the objectivity of the scoring rubric, while noting an aversion to the consequences of scoring negatively:

“it can explain [to] me what [they] can do and what I should pay attention for. And I’m just got to follow... the suggestion... and I try to be more objective than subjective because sometimes it’s very difficult to tell how good the resident is, because all of them, they look very good to me. For it was very difficult to put like a negative mark, but sometimes, you know, like when they cannot find the picture, I push myself to put something negative. But most of the time they were so good and I was very impressed. And I’m just going to follow the [ECAT] paper and just follow the mark[ing scheme] and I think it’s helpful.”

The rater felt the tool was a guide that made assessment easier in a time-sensitive situation where the person doing the scan was stressed and “trying to be faster than me.” Rater 2 self-identified as a direct observer of the assessment process, and as a sonographer, which came into play in the observation:

“From [a] sonographer point of view, it was very, very nice to see how the young doctors, how they deeply understand what they're doing and how they're struggling to be the best with the picture. Because honestly, it is not their job, it is our job – like it is something extra. And from a sonographer point of view, I think for us, it’s very nice to have a doctor – the cardiologist who is working with you, who really understands how can you make a picture. How can, you know, the struggling to do the picture.”

Rater 2 offered some constructive feedback to improve the rating form in rater training for ease-of-use by raters, with more explanation of how the global assessment of diagnostic quality is determined, and adding a section for raters to comment on what they felt the participants did best.

Rater 3 articulated personal challenges in using the ECAT to provide feedback, and the construct they felt the tool was assessing, as they perceived that formative assessment and the skills assessed were best done in a clinical setting on a real patient. Rater 3 reportedly did not use the free-text box to provide specific narrative feedback because he/she felt limited by not being able to see the trainee performance: “I guess because you couldn’t really tell where they messed up. You couldn’t see when they were doing their images how, what they struggled through or what they didn’t.” Rater 3 understood the study design rationale for simulator standardization and assessing a focused scope of skills by the tool, but queried if this impeded useful feedback, stating: “just looking at the video clip and saying whether they utilize the screen depth and width accurately is a little too little for me, to decide whether they’re actually able to get that image.” Rater 3 also raised additional concerns that the verbal anchors may not have represented trainee knowledge, saying, “so there were some people who would score high or low on that and I got the sense they either, they were scoring low but they knew more. Or they scored high but they didn’t actually know it.” At the same time, rater 3 stated that if trainees were aware of the assessment tool, they would “train for the test,” which would “discriminate those who had access to the list of desired skills”.

Another issue raised by rater 3 was that immediate, not delayed, feedback was imperative for training in psychomotor skills. Rater 3 did not feel there was benefit in providing offline comments for the skill-sets assessed, saying: “it has to be real-time feedback, not recorded feedback,” and when asked to comment on participant concerns about barriers to real-time feedback, stated that perhaps it was personality or temperament related. Rater 3 proposed that feedback could have been improved if raters were given the opportunity to directly watch trainees obtain pictures on real patients to allow them to exercise techniques to optimize images not possible on simulators, if trainees could interpret “canned” images rather than self-acquired images on the ECAT, ensuring all required structures on complementary views were identical, and improving the audio quality and identification aid.

There were disparate findings for raters’ perceived value of the ECAT testing and feedback. One rater highlighted aspects raised in participant feedback such as optimal timing of feedback delivery, and juxtaposed his/her comments alongside standard teaching as an educator. Two raters remarked on the ease of use of ECAT testing and feedback provision, emphasizing the scoring rubric was more objective than the standard, which made it easier to provide concrete, constructive feedback to help participant learning plans. However, one rater felt limited in using the tool, and had concerns regarding the tool’s aligned anchors, the timing of feedback, and the role of formative feedback. This rater’s perceptions of how participants optimally learn were occasionally contrary to the reported participant perceptions. The heterogeneity in rater response highlights some strengths and areas for improvement when interpreting implications inferences from the rater perspective.

6.3 Evidence for Scoring

We explored whether raters used the ECAT consistently when scoring by calculating the intra-class correlation coefficient (ICC) for all four raters. Across all items, the ICC was 0.837 (95% CI 0.81-0.87). For ECAT score, the ICC was 0.913 (95% CI 0.81-0.97). The magnitude of ICC is considered good for formative assessment.

Participants' scores ranged across each of the two dimensions assessed for the eleven echocardiography views. We report the mean score (standard deviation) for each item, where the minimum potential score was 0 and maximum potential score was 3 (Table IV).

TABLE IV – TEST ITEM DESCRIPTIVE STATISTICS

Variable	Mean (Standard Deviation)
Apical two chamber window use of screen	0.9 (0.8)
Apical three chamber window use of screen	1.0 (1.0)
Apical five chamber window use of screen	1.1 (0.9)
Parasternal short axis mitral valve window use of screen	1.1 (1.1)
Parasternal long axis window use of screen	1.2 (0.8)
Parasternal short axis aortic valve window use of screen	1.2 (0.8)
Apical four chamber window use of screen	1.2 (0.9)
Parasternal short axis papillary muscle window use of screen	1.2 (1.0)
Right ventricular inflow window use of screen	1.2 (1.1)
Apical five chamber window anatomy identification	1.3 (0.6)
Right ventricular inflow window anatomy identification	1.3 (1.0)
Apical three chamber window anatomy identification	1.3 (1.0)
Parasternal short axis apex window use of screen	1.3 (1.2)
Apical four chamber window anatomy identification	1.5 (0.6)
Parasternal short axis aortic valve window anatomy identification	1.5 (1.1)
Subcostal window use of screen	1.5 (1.1)
Apical two chamber window anatomy identification	1.7 (0.9)
Parasternal short axis mitral valve window anatomy identification	1.9 (0.9)
Subcostal window anatomy identification	2 (0.8)
Parasternal long axis anatomy identification	2 (0.9)
Parasternal short axis papillary muscle window anatomy identification	2.0 (0.9)
Parasternal short axis apex window anatomy identification	2.6 (1.0)

Our exploratory factor analysis revealed a two-factor model, which we labeled “use of screen” and “identification of anatomy” ([Table V](#)). Inflections in plotted Eigenvalues suggested the model was slightly improved with three factors, but without clear delineation of which dimensions to include, we retained the two-factor model.

TABLE V: EXPLORATORY ROTATED FACTOR PATTERN ANALYSIS

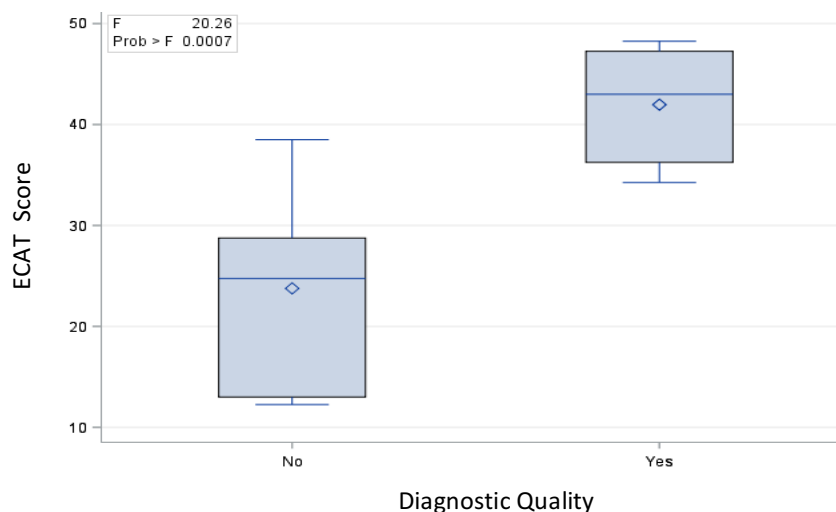
Rotated Factor Pattern	Factor 1	Factor 2
Parasternal short axis papillary muscle use of screen	0.88	0.18
Apical five chamber use of screen	0.87	0.28
Parasternal short axis apex use of screen	0.86	-0.08
Parasternal short axis mitral valve use of screen	0.85	0.29
Parasternal short axis aortic valve use of screen	0.83	0.07
Apical two chamber use of screen	0.82	0.18
Apical four chamber use of screen	0.80	0.30
Subcostal use of screen	0.79	0.24
Apical three chamber use of screen	0.78	0.16
Right ventricular inflow use of screen	0.74	0.46
Parasternal long axis use of screen	0.72	0.33
Parasternal short axis apex anatomy identification	0.34	0.27
Parasternal short axis aortic valve anatomy identification	0.28	0.82
Parasternal long axis anatomy identification	0.16	0.76
Subcostal anatomy identification	0.16	0.70
Parasternal short axis mitral valve anatomy identification	0.39	0.69
Apical four chamber anatomy identification	0.21	0.65
Parasternal short axis papillary muscle anatomy identification	0.11	0.61
Apical three chamber anatomy identification	0.21	0.60
Right ventricular inflow anatomy identification	0.15	0.59
Apical five chamber anatomy identification	-0.11	0.55
Apical two chamber anatomy identification	0.26	0.42

We also computed coefficient alpha to determine the internal consistency of the items assigned to the two-factors in the model. For the dimension, “use of screen coefficient alpha was 0.96, and the dimension, “identification of anatomy,” coefficient alpha was 0.87. These findings suggest the items within each factor were highly related.

6.4 Evidence for Extrapolation

We explored the association between ECAT score and diagnostic quality using a generalized estimating equation model and linear model. We modelled “diagnostic quality” as a yes/no variable to account for the ordinality of the item-level scores. Both models suggested a significant difference in the average ECAT score when the diagnostic quality was judged as ‘yes’ vs. ‘no’, with higher scores associated with ‘yes’. Using the linear model, F-statistic was 20.26 ($p=0.0007$) (Figure 2).

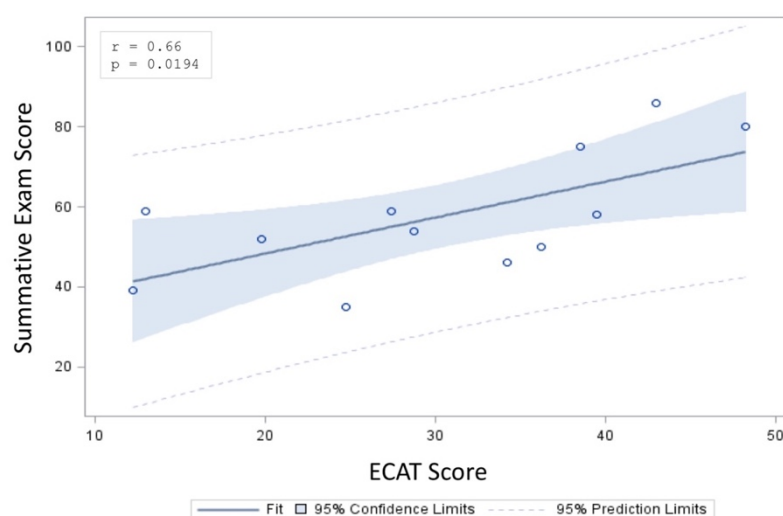
FIGURE 2: Association between ECAT score and diagnostic quality assessment



The correlation between ECAT score and summative testing is shown in Figure 3. Data for summative end of year echocardiography examination was available for 12 participants of 14,

with two postgraduate fellows excluded as summative in-training examinations are only offered to cardiology trainees during their residency. The correlation between ECAT score and end of year examination was statistically significant ($r=0.66$, $p=0.02$).

FIGURE 3: Correlation between ECAT score and summative exam score



Correlation between ECAT score and self-reported variables is shown in **Table VI**. There was a significant correlation between ECAT performance and number of echocardiograms seen ($r=0.64$, $p=0.014$), number of echocardiograms performed ($r=0.65$, $p=0.012$), and number of echocardiograms interpreted ($r=0.60$, $p=0.024$). Variables which were not discernibly associated included number of ultrasounds seen, performed, and interpreted. Data was distributed parametrically.

TABLE VI – CORRELATION BETWEEN ECAT SCORE AND OTHER VARIABLES

Self-reported Variables	Coefficient	p-value
Number of previous echocardiograms seen	0.64	0.014
Number of previous complete echocardiograms performed	0.65	0.012
Number of previous complete echocardiograms directly interpreted	0.60	0.024
Number of previous ultrasounds seen	0.46	0.095
Number of previous ultrasounds performed	0.18	0.55
Number of previous ultrasounds directly interpreted	0.01	0.98

ECAT score increased according to level of training ([Figure 4](#)). We conducted a one-way ANOVA to compare ECAT scores across three trainee groups, defined as year 1, 2, ≥ 3 , which was significant at $p=0.01$. We used the least significant difference post-hoc test to conduct pairwise comparisons of groups. We found significant differences between 1st years vs 2nd years and 1st years vs $\geq 3^{\text{rd}}$ years, but no significant difference between 2nd years and $\geq 3^{\text{rd}}$ years ([Table VII](#)).

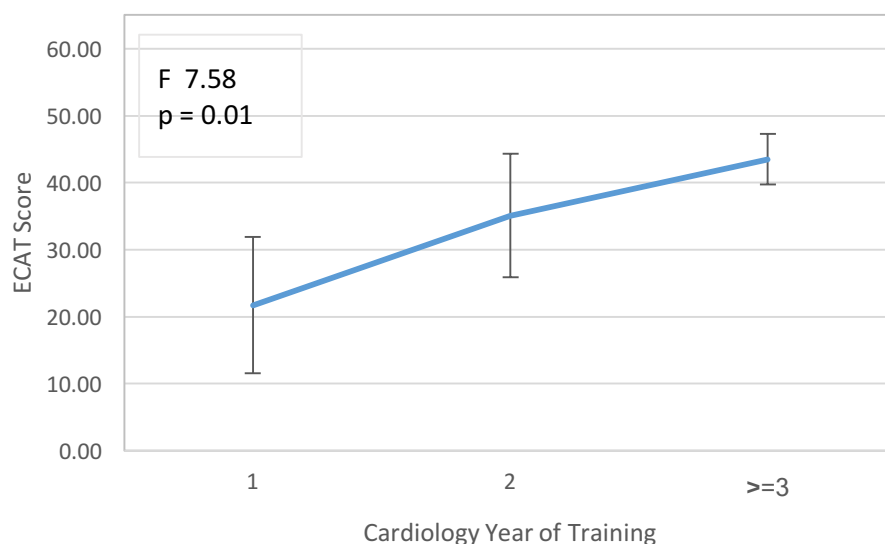
FIGURE 4: ECAT Score according to level of training

TABLE VII: Pairwise comparisons between participant groups

Comparison	p-value
1 st year vs 2 nd year	0.03
1 st year vs $\geq 3^{\text{rd}}$ year	0.003
2 nd year vs $\geq 3^{\text{rd}}$ year	0.17

7: DISCUSSION

7.1 Validity Argument: Judgment on our Interpretation-use Argument

We collected validity evidence to evaluate the use of an ECAT for formative assessment of cardiology trainees' simulation-based basic echocardiographic skills. We applied Kane's validity framework by specifying an interpretation-use argument, stating assumptions and hypotheses for various inferences, and creating an assessment plan to test our claims through qualitative and quantitative analyses. Below, we synthesize the evidence accrued for the implications, scoring and extrapolation inferences, and provide a judgment on our interpretation use argument: whether to accept, reject, or revise our initial interpretation-use argument and/or use of the proposed assessment tool.

7.1.1 The Implications Inference

For the implications inference, we consider how our findings support our claim that participants and raters perceived the ECAT testing and feedback facilitated learning and improved echocardiography skills.

We made three decisions at study inception which increased the yield of validity evidence for the implications inference, and distinguished our study from others in the reviewed literature. First, we prioritized hypotheses for implications validity evidence. Second, we designed the study to capture narrative data reflecting the perceptions of participants and raters, seeking a deeper understanding of how meaningful the assessment was to those using it. Third, we studied the

impact of a formative, rather than summative, assessment, as feedback might be more effectively used by learners in this context [101].

The majority of participants reported that feedback stimulated personal change, both in concrete action and in perceived self-assessment. Both results are consistent with what is known about feedback in simulation-based medical education – that information about previous performance intended to guide future performance is a key feature [32, 102, 103]. Change was unique to each individual, ranging from a participant making physical changes in the angle of their hand to generate an echocardiographic window, to being motivated by his/her feedback to look up how to approach visualizing a specific structure. Training level seemed to matter, with more senior trainees reflecting that if they were at an earlier stage in their career, different aspects of feedback would have been more helpful, and they would have used it differently. Findings are attenuated somewhat by comments from two individuals who felt feedback received was useful, but when asked to elaborate, either did not recall specific concrete changes made based on this feedback, or felt they did not have the background knowledge to interpret the feedback. This finding recalls work by Humphrey-Murto et al. in a study of resident recall immediately after an OSCE and one month later, where they found residents recalled very few feedback points at both times, and what was recalled was neither very accurate nor representative of the feedback actually provided [104].

How feedback was delivered impacted participants' perceived learning. Multiple modes of feedback were preferred, and participants felt this offered an opportunity to review and

consolidate their learning. This finding aligns with the meta-analysis by Hatala et al. examining the effectiveness of feedback in simulation-based medical education, where multiple sources of feedback lead to enhanced learning outcomes [105]. Participants' response to timing of feedback was variable, with many citing their lack of ability to have direct dialogue with those providing feedback as a concern. There was tension between wishing to receive immediate vs. delayed feedback. In interpreting this finding, we consider that Hatala et al. found for novices, terminal feedback was more effective than concurrent feedback for skills retention [105]. To explain this, they considered theories such as the "guidance hypothesis" where learners provided with concurrent feedback may form a reliance on the feedback rather than developing skill through independent "struggle time" [106]. Hatala et al. also considered cognitive load theory, where the amount of information provided through feedback might best be balanced by an instructor, who can make a task more understandable [105]. In our study, the majority of participants felt they benefited from being able to access and review their recorded performance data and written feedback, but some participants and one of the raters believed that immediate feedback was required for psychomotor skills development.

Participants' judgments of the credibility of the assessment influenced their receptivity to feedback. Credibility of feedback providers is known to impact the degree to which students accept feedback [107]. In our study, participants knew that expert raters were directly observing their performance either in person or by video, and most judged their feedback to be credible. Multiple expert raters representing two different professions (echocardiographers and sonographers), provided feedback, but as this was only revealed post-hoc, its impact could not

be interpreted. Participants felt the assessment overall was fair, ascribing credibility to the process. Participants outlined some pros and cons to credibility of the simulator with which they were assessed, which some felt helped and others felt hindered their learning. One of the pros was that the echocardiography simulator seemed quite realistic compared to doing an echocardiogram: the probe felt the same way, the simulated echocardiogram image on the screen looked like an image that would be achieved in a real-life setting. However, realism was not perfect: participants noted the simulator could not be positioned in order to acquire better images, scanning either yielded a perfect image or no image, with no grey zone in between as seen in real life, and some views were impossible to achieve on the simulator such as the suprasternal window. Dieckmann [108] stated that “there should be enough realism of the right type for the purpose in a way that participants see the simulation endeavor as a relevant environment for the[ir] goals”. Balancing participants comments on pros and cons of their perceived realism of the simulator reminds us of Dieckmann’s equation: “Simulation = (Reality – X) + Y, where X are departures from reality, and Y are affordances gained through the simulation experience.” In our study, participant-identified deviations of the simulator from clinical reality (“X”) may have been offset by affordances that participants’ identified as an enhancement to their clinical teaching (“Y”), for example, dynamic conceptualization of visuospatial planes, and anatomic labelling in the patient avatar [109]. From the cumulative commentary, it is difficult to say whether the net balance of the reality and deviations from reality ultimately ended up as a net benefit to participant learning. Considering all aspects of the assessment process as a whole, however, overall credibility was considered reasonable.

In conclusion, the themes identified in participant and rater interviews include: feedback stimulated change, how feedback was delivered impacted participants' perceived learning, and assessment credibility influenced participants' reception of feedback, and generally the results supported our claim that participants and raters perceived the ECAT testing and feedback facilitated learning and improved echocardiography skills. A minority of participants and raters discussed dissenting views of their recall of feedback, which does attenuate the strength of evidence somewhat, as if individuals can not recall feedback, it would be challenging to make the case that it facilitated learning or improved skills. Due to the small sample size, the relative significance of this compared to the majority response from the data, is uncertain.

There were some nuances in rater and participant perception of the optimal timing of assessment and the credibility of the simulator, but we feel this does not affect the overall strength of the evidence and provides depth to the discussion, signalling future areas to explore.

7.1.2 The Scoring Inference

For the scoring inference, we consider how our findings supported our claim that raters would use the ECAT consistently when scoring, and that dimensions on the ECAT were unique and demonstrated evidence of independence.

Evidence in our literature review was generally favorable for the scoring inference: most scoring rubrics captured key aspects of performance, and were used fairly and accurately. In our study design, we incorporated elements into the ECAT based on assessment tools with the highest quality scoring validity evidence. For example, we aligned the functional elements of the task

with professional standards refined by content experts, we adapted a previously studied instrument, we required rater training, and employed methods such as de-identifying videos for blinded analysis and ensuring real-time raters had never interacted with participants clinically, to reduce rater bias. Our study was unique compared to others in the literature, as our results yielded both quantitative and qualitative scoring validity evidence, which supported our interpretation-use argument.

There was high internal consistency for the items on the ECAT. Some may consider that this consistency suggests the tool measures only one performance dimension, however when we conducted a factor analysis, the model suggested two factors, for which constituent items were scored consistently. The observed inter-rater reliability was good, and in a range appropriate for a formative assessment.

In conclusion, we believe our data favorably support our hypotheses about scoring. Raters appeared to use the ECAT consistently when scoring and dimensions on the ECAT were unique and demonstrated evidence of independence.

7.1.3 The Extrapolation Inference

For the extrapolation inference, we consider how our findings supported our claim that participants' ECAT scores would be positively associated with observed performance on end-of-year echocardiography examination, were positively related to the global impression of whether

the simulated echocardiogram is of diagnostic quality, and would discriminate trainees according to expected level of performance by their postgraduate training year.

In our interpretation-use argument, extrapolation validity evidence was a secondary focus to the other types of validity evidence sought above. This is because, while frequently cited, likely due to a clinical bias when it comes to assessment: many clinical educators and residency programs need to know that their assessment correlates with a clinically meaningful endpoint, correlations used to bolster this claim may be due to myriad factors beyond an education or training effect. We observed a positive correlation between level of training and assessment score which provides proof of principle that individuals in higher training years achieved higher scores and vice versa, but as Cook and Hatala state, observed correlations between expertise and score might arise from factors which might be unrelated to the intended construct, and is considered weak evidence [44] . This finding does responds to a rater comment in which the rater was skeptical that test performance on our tool was related to level of expertise. Thus, this evidence was weakly supportive of our validity argument.

We demonstrated correlation between assessment score and a criterion-referenced measure, an end-of-year summative echocardiography examination. The association between a criterion-based finding results in more compelling validity evidence; the challenge many researchers face is trying to find a plausible criterion-based metric for which to seek an association. We were able to access data from the residency program end-of-year examination, a purely cognitive assessment of interpretive echocardiography skills. The demonstrated association aligns with

previous work in which Nair et al. postulated a direct relationship between echocardiography psychomotor skills and cognitive interpretive skills [30]. Our criterion-based evidence is more discriminating in evaluating validity claims than simple expertise-performance correlations, and supports the validity argument.

Another clinically relevant metric for which we sought a correlation between ECAT scores was whether raters deemed a complete echocardiogram of diagnostic quality. This outcome is analogous to a global rating score, as it provides a global determination of the participant's overall performance. In the cardiology world, where trainees may scan real patients asynchronously and images are analyzed by cardiologists to make clinical management decisions, clinical care would be impaired if images were not of sufficient diagnostic quality. This may lead to, at best, the patient having to come back to the echocardiography laboratory for additional imaging, and at worst, missed or incorrect diagnoses and management. The ECAT score was highly correlated with whether an image was of diagnostic quality. This provides validity evidence to support our claim.

In conclusion, we believe our data support our extrapolation inferences. While some extrapolation validity evidence was more favorable than others, we did provide at least two clinically plausible associations in addition to the usual association between expertise and score, which helps demonstrate that the ECAT is associated with relevant outcomes in the real world.

7.1.4 The Generalization Inference

Our interpretation-use argument and hypotheses did not specify generalization inferences: that the items sampled in the ECAT represented all theoretically possible items sought in a complete echocardiogram. This is because the ECAT represented the comprehensive skill set all cardiology trainees must achieve, instead of a subset of views seen in more rapid protocols used in critical care, anesthesia and emergency medicine [37, 61, 67]. Sampling was thus less important than trying to achieve a complete exam. In describing his framework, Kane notes that not all four inferences need necessarily be addressed, and those inferences specified are a function of the interpretation-use argument [1]. Bordage notes that “conceptual frameworks are dynamic entities, and benefit from being challenged and altered as needed,” which is what we have done with respect to the generalization inference [110]. Thus, in our study, we can not comment on generalization validity evidence to support or refute our validity argument.

7.1.5 Other Factors Impacting the Validity Argument

Beyond appraising validity evidence, an important and understudied aspect in validation research is feasibility in development, implementation and interpretation of scores. Money and human resource issues play a key role in wider implementation and scalability of initial pilot projects. In designing this project, funding and feasibility were significant challenges which we had to address in multiple iterations of study design. Our study contained costs by seeking curricular support from an academic fund, limiting the protocol to a single test administration, recruiting volunteer raters, acquiring access to the simulator for a nominal rate through divisional collaboration, and managing the project internally rather than through a research

coordinator. We would have wished to have multiple test administrations as per the suggested multiple frequent low-stakes assessments recommended by ACGME Milestones, but faced logistical and political challenges in trying to implement this due to the perception of interrupting scheduled echocardiography learning rotations.

7.1.6 Summary Judgment

We argue that collectively, the ECAT appears to be largely supported as a tool for formative assessment of novice trainees practicing simulation-based basic echocardiography scanning and interpretation. The bulk of evidence collected in the current study supports the implications, scoring, and extrapolation inferences. Our validity argument is not an end in and of itself, but part of a constructive process as we look ahead to establish and refine validity evidence for the ECAT in simulation-based assessment in future studies.

7.2 Impact on Future Research

This study, the first to systematically consider and generate validity evidence for a formative assessment of simulation-based basic echocardiography skills, can provide practical insights for echocardiography educators, medical educators and researchers. From an echocardiography perspective, this addresses a key need identified in the updated CoCATS 4 Task Force 5 guidelines for training in echocardiography [24], which suggests, but does not specify the exact nature of, multiple assessments and simulation as learning modalities for trainees. We have devised a robust tool for use in formative assessment, which can address these needs. Cardiology training programs may incorporate multiple formative assessments based on this tool

into their longitudinal curricula as residents progress through the training program. There may be an opportunity to incorporate the ECAT into assessments of competence through ACGME Milestones or CanMEDS 2015 Physician Competency Frameworks [10, 15].

Medical educators may use the process of seeking validity evidence presented here as a template for devising assessments in their own medical specialties. The approach of our study, using validity frameworks to analyze assessment testing, builds on the work of others who have sought to utilize validity frameworks in evaluating various constructs [54-59, 109]. One important lesson learned through this study was the benefit of qualitative evidence, which yielded insights beyond the typical implications evidence seen in the literature to date, and could be explored further. An example of one such insight was that the majority of participants felt they benefited from being able to access and review their recorded performance data and written feedback, but some participants and one of the raters believed that immediate feedback was required for psychomotor skills development – the tension between these perspectives could be explored in future studies. Emergent findings from open-ended feedback could be used constructively to target future educational interventions.

Future avenues for research were identified through the process of collecting and analyzing validity evidence in our study and in the context of the wider assessment literature. From the conceptualization of the interpretation-use argument, we realized that if our tool were used for a summative, rather than formative purpose, different data would need to be acquired, as, according to Cook and Hatala, “validity evidence applies only to the purpose, context and learner

group in which it was collected” [44]. If in future, the purpose of the tool was expanded to a summative assessment, a larger sample and clearly outlined generalization evidence could be helpful, which might include generalizability studies. Future research to obtain stronger scoring validity evidence could emphasize rigorous rater training and expand upon the in-depth qualitative collection we attempted to do in our study to capture rater and participant feedback on the assessment and the overall educational experience. To further advance extrapolation validity evidence, studies could look for correlations comparing plausible connection between the ECAT and echo outcomes in the real world. It could be valuable for future research to consider the use of the “cumulative sum of scores” or time-based training effects in a time-series analysis to monitor formative progress [79], and think about how to establish the ideal length and structure of educational assessments [56] using validity frameworks. Implications validity evidence could be sought in studies considering simulation-based assessment as a way to alleviate human resource reliance in training programs [77,87] and explicitly incorporating the consequences of assessment on educators and learners into study aims [61], as we strived to do in our study.

7.3 Limitations

The small sample size of our study resulted in reduced statistical power for some of our analyses. However, this study can be considered an exploratory, preliminary step towards larger data generation, with an emphasis on feasibility. We approached all core trainees in our large, single-center cardiology training cohort, but as testing occurred after-hours during summer months, only a subset could participate. Our sample included over half of the trainees, with excellent

representation of different levels of expertise. Reviewing data for workplace-based assessment and educational studies, our sample size was considered within the acceptable range. Although the inferences drawn from our small sample size are mostly hypothesis generating, and our reliability estimates limited, the results of this study demonstrated the ECAT was reliable, consistent, able to distinguish between groups of expertise, correlated with summative testing, and was perceived to be useful in facilitating the feedback process.

Although the study did not entail any material risks, participants may have felt judged by rater perceptions regarding their echocardiography competency. The study was designed to minimize the perception of being judged by ensuring that all participants received a unique identifying number tracking their performance through the study, with de-identified assessments and feedback. Real-time assessment was only performed by rater(s) who did not have direct involvement in trainee education and evaluation, and faculty evaluating off-line performance were blinded to participant identity. Participants were reassured that all results would remain confidential and not shared with their Program Director or clinical supervisors.

We formulated an interpretation-use argument and collected validity evidence solely for formative assessment of basic technical and evaluative skills of cardiology trainees. Should the ECAT be considered for another purpose, such as summative use as an end-of-year examination, this evidence may not be defensible, and different validity evidence may need to be collected and considered for this purpose.

In addition to being purpose-specific, assessment tools are context-specific. The proposed tool was designed to assess basic echocardiography skills in a simulation-based, standardized context. We made thoughtful changes to the tool, articulating changes clearly in the methodology and ensuring raters were aware of this in rater training. If ECAT testing is proposed for use with clinical patients using echocardiography machines, validity evidence may not be sufficient and alternative validity evidence may need to be collected and considered.

While conducting this study, we experienced challenges in how to prioritize reporting and collecting evidence for Kane's four inferences. Our interpretation-use argument was formulated around implications, scoring and extrapolation evidence, but we were unsure as to whether to include or exclude the generalization inference. We decided to not specify generalization inferences in our interpretation-use argument and hypotheses as upon considering our test, it was intended to be representative of the required skill set all cardiology trainees must achieve through a formative process, and we were not seeking to sample the best subset of views to perform a focused, partial, echocardiogram. If in future, the purpose of the tool was expanded to a summative assessment, a larger sample and clearly outlined generalization evidence could be helpful.

We used a content analysis framework for the qualitative component of our study. There were pros and cons to this approach. We thought it would be helpful as our study design was seeking to describe a phenomenon, which is the context in which content analysis is usually used. We did not have any theoretical perspectives we imposed on our acquisition of data. However we

acknowledge that this may have limited development of a complete understanding of the context, and be more rudimentary than the deeper understanding acquired through a grounded theory or phenomenology framework. A future study using a more comprehensive qualitative analysis framework may be able to provide further insights into the role of feedback in formative assessment of simulation-based basic echocardiography skills.

7: CONCLUSIONS

Our study is the first project to systematically consider and generate validity evidence for the ECAT, which is designed to assess simulation-based basic echocardiography skills for formative assessment of trainees. Kane's argument-based validity framework provided a method to prioritize, synthesize, and evaluate evidence in our study. Our results found the ECAT to be a valued tool which both trainees and raters report stimulated a feedback process they perceived would help with goal setting, which demonstrated good inter-rater reliability, and which correlated well with level of expertise and a summative assessment. Our approach builds on appeals from researchers in medical education to utilize validity frameworks in evaluating assessment tools and assessment programs [54-59, 109]. Our findings provide practical insights for echocardiography educators and for those in medical education. The process of evaluating the appropriateness of the interpretations, uses and decisions stemming from assessment results will assist in future development of instruments for competency-based assessments in cardiology and beyond.

CITED LITERATURE

- 1) Kane MT. Validation. In: Brennan RL (ed). Educational Measurement. 4th ed. Westport: Praeger, 2006:17-64.
- 2) Frenk J, Chen L, Bhutta Z et al. Health Professionals for a new century: transforming education to strengthen health systems in a interdependent world. *Lancet* 2010; 376: 1923-58.
- 3) Frank JR, Snell LS, ten Cate O et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8):638–645.
- 4) Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002; 287: 226–35.
- 5) Gruppen LD, Mangrulkar RS, Kolars JC. Competency-based education in the health professions: implications for improving global health. Commission paper 2010. [http://www. globalcommehp.com/](http://www.globalcommehp.com/) (accessed Oct 30, 2016).
- 6) Sullivan RL. 1995. *The Competency-Based Strategy Paper No 1*. JHPIEGO Corporation: Baltimore, Maryland.
- 7) Carraccio C, Wolfsthal SD, Englander R et al. Shifting paradigms: From Flexner to competencies. *Acad Med* 2002; 77(5):361–367.
- 8) Harden RM, Crosby JR, Davis MH et al. AMEE Guide no. 14: Outcome-based education: Part 5 – from competency to meta- competency: A model for the specification of learning outcomes. *Med Teach* 1999; 21(6):546–552.
- 9) Pugh D, Regehr G. Taking the sting out of assessment: is there a role for progress testing? *Med Educ* 2016; 50: 721-729.
- 10) Griffin S, Flaherty J (chairs), et al. The Internal Medicine Subspecialty Milestones Project. (2015). Retrieved from <https://www.acgme.org/acgmeweb/Portals/0/PDFs/Milestones/InternalMedicineSubspecialtyMilestones.pdf>
- 11) Caverzagie KJ, Iobst WF, Aagaard EM, et al. The Internal Medicine Reporting Milestones and the Next Accreditation System. *Ann Intern Med*. 2013; 158:557-9.
- 12) Green ML, Aagaard EM, Caverzagie KJ, Chick DA, Holmboe E, Kane G, et al. Charting the road to competence: developmental milestones for internal medicine residency training. *J Grad Med Educ*. 2009;1(1):5–20.

- 13) Frank J, Snell L, Sherbino J (2014). Draft CanMEDS 2015 Physician Competency Framework.
Retrieved from
http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/canmeds2015_framework_series_II_e.pdf
- 14) Rubin P, Franchi-Christopher D. New edition of tomorrow's doctors. *Med Teach*. 2002;24(4):368–369.
- 15) Simpson JG, Furnace J, Crosby J et al. The Scottish doctor—learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. *Med Teach*. 2002;24(2):136–143.
- 16) Norman G, Norcini J, Bordage G. Competency-based Education: Milestones or Millstones. *J Grad Med Ed* 2014. 1-6.
- 17) Sherbino J, Kulsegaram K, Worster A, Norman GR. The reliability of encounter cards to assess the CanMEDS roles. *Adv Health Sci Educ Theory Pract*. 2013;18(5):987–996.
- 18) Kassam A, Donnon T, Rigby I. Validity and reliability of an in-training evaluation report to measure the CanMEDS Roles in emergency medicine residents. *CJEM*. 2013;15(0):1–7.
- 19) Yu E, Nair P, Sibbald M et al. Can diagnostic and procedural skills required to practice cardiology as a specialist be mastered in 3 years. *Can J Cardiol* 2014; 31: 91-94.
- 20) Watson DR, Flesher TD, Ruiz O, Chung JS. Impact of the 80-hour workweek on surgical case exposure within a General Surgery residency program. *J Surg Educ*. 2010;67(5):283-289.
- 21) Sadaba JR, Urso S. Does the introduction of duty- hour restriction in the United States negatively affect the operative volume of surgical trainees? *Interact Cardiovasc Thorac Surg*. 2011;13(3):316-319.
- 22) Peets A, Ayas NT. Restricting resident work hours: the good, the bad, and the ugly. *Crit Care Med* 2012;40:960-6.
- 23) Palaniswamy C. Milestones and the Next Accreditation System. *J Am Coll Cardiol* 2014; 64(11): 1178-1180.
- 24) Ryan T (chair). CoCATS 4 Task Force 5: Training in Echocardiography. Endorsed by the American Society of Echocardiography. *J Am Coll Cardiol*. 2015; 65(17): 1786-1799.

- 25) Rambihar S, Khanduja KK, Nesbitt GN et al. Use of the Modified Delphi Technique to Develop a Competency-Based Simulation-Enhanced Transthoracic Echocardiography Curriculum for Novices. *J Am Soc Echocardiography* 2014; 27(6): B97.
- 26) Feigenbaum H. Educational problems in echocardiography. *Am J Cardiol* 1974; 34: 741-2.
- 27) Burwash IG, Basmadjian A, Bewick D et al. 2010 Canadian Cardiovascular Society/ Canadian Society of Echocardiography Guidelines for Training and Maintenance of Competency in Adult Echocardiography. *Can J Cardiol*. 2011; 27(6):862-4.
- 28) Ryan T. CoCATS Task Force 4: Training in Echocardiography. Endorsed by the American Society of Echocardiography. *J Am Coll Cardiol*. 2008; 51(3): 36-42.
- 29) Yu E. The assessment of technical skills in a cardiology training program: is the ITER sufficient? *Can J Cardiol* 2000; 16(4):457–462.
- 30) Nair P, Siu S, Sloggett C et al. The Assessment of Technical and Interpretive Proficiency in Echocardiography. *J Am Soc Echocardiography* 2006; 19(7): 924-931.
- 31) National Board of Echocardiography 2016. Accessed Nov 1, 2016: <http://echobords.org/content/ascexam®>
- 32) Issenberg, S.B., McGaghie WC, Petrusa E et al. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Educ* 2005, 27(1): 10–28.
- 33) Motola I, Devine L, Chung HS, et al. Simulation in healthcare education: A best evidence practical guide, AMEE Guide No. 82 *Med Teach* 2013; 35: e1511-e1530.
- 34) Abrahamson S, Denson JS, Wolf RM. Effectiveness of a simulator in training anesthesiology residents. *J Med Educ*. 1969; 44: 515-519.
- 35) Montealegre-Gallegos M, Mahmood F, Kim H et al. Imaging skills for transthoracic echocardiography in cardiology fellows: the value of motion metrics. *Ann Cardiac Anes* 2016; 19(2): 245-250.
- 36) Matyal R, Bose R, Warraich H et al. Transthoracic Echocardiographic Simulator: Normal and the Abnormal. *J Cardiothorac Vasc Anes* 2011; 25(1):177-181.
- 37) Neelankavil J, Howard-Quijano K, Hsieh TC, et al. Transthoracic echocardiography simulation is an efficient method to train anesthesiologists in basic transthoracic echocardiography skills. *Anesth Analg*. 2012 Nov;115(5):1042-511.

- 38) Matyal R, Motealegre-Gallegos M, Mitchell J et al. Manual Skill Acquisition During Transesophageal Echocardiography Simulator Training of Cardiology Fellows: A Kinematic Assessment. *J Cardiothoracic Vasc Anes* 2015; 29(6): 1504-1510.
- 39) Mahmood, F. Training in echocardiography – Top-Down or a Bottom-Up Approach? *J Am Soc Echocardiography* 2014; 27(10): 18A-19A.
- 40) Cook DA, Brydges R, Ginsburg S et al. A contemporary approach to validity arguments: a practice guide to Kane’s framework. *Med Educ*. 2015 Jun; 49(6):560-75.
- 41) Cook DA, Lineberry MJ. Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Acad Med*. 2016 Jun; 91(6):785-795.
- 42) Downing SM. Validity: On meaningful interpretation of assessment data. *Med Educ*. 2003;37:830–837.
- 43) Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd ed. New York, NY: American Council on Education and Macmillan; 1989:13–103.
- 44) Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv in Sim* 2016; 1(31): 1-12.
- 45) Cureton, E. E.. Validity. In E. F. Lindquist (Ed.), *Educational measurement* 1951: pp. 621-694. Washington, DC: American Council on Education.
- 46) Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). 1971. Washington, DC: American Council on Education.
- 47) Association AER. American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 2014.
- 48) Downing S and Haladyna S. Validity and its threats. Chapter 2 in *Assessments in Health Professions Education* (Ed. Yudkowsky and Downing) 2009 Taylor and Francis: London and NY.
- 49) Linn RL. Evaluating the validity of assessments: The consequences of use. *Educ Meas Issues Pract*. 1997;16:14–16.
- 50) Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;19:166.e7–16.

- 51) Cronbach, L. J. Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). 1988 Hillsdale, NJ: Lawrence Erlbaum.
- 52) Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas* 2013;50:1–73.
- 53) Clauser B, Margolis M, Holtman M et al. Validity considerations in the assessment of professionalism. *Adv in Health Sci Educ* 2012; 17:165-181.
- 54) Cook, DA, Zendejas B, Hamstra S et al. What counts as validity evidence? *Adv Health Sci Edu* 2014; 19(2): 233-50.
- 55) Hawkins R, Margolis M, Durning S et al. Constructing a validity argument for the mini-clinical evaluation exercise: a review of the research. *Acad Med* 2010; 85: 1463-61.
- 56) Boulet J, Jeffries P, Hatala R et al. Research regarding methods of assessing learning outcomes. *Simul Healthcare* 2011; 6:S48-51.
- 57) Hatala R, Cook DA, Brydges R et al. Constructing a validity argument for the objectives structured assessment of technical skills (OSATS): a systemic review of validity evidence. *Adv in Health Sci Educ* 2015.
- 58) Kreiter C, Otaki J. Constructing a more comprehensive validity argument for medical school admission testing: predicting long- term outcomes. *Teach Learn Med* 2015; 27(20): 197-200.
- 59) Till H, Ker J, Myford C, et a. Constructing and evaluating a validity argument for the final-year ward simulation exercise. *Adv in Health Sci Educ* 2015; 1263-1289.
- 60) Nielsen D, Gotzsche O, Eika B. Objective Structured Assessment of Technical Competence in Transthoracic Echocardiography: A Validity Study in a Standardized Setting. *BMC Medical Education* 2013; 13:47.
- 61) Millington S, Arntfield R, Hewak M, et al. The Rapid Assessment of Competency in Echocardiography Scale. *J Ultrasound Med* 2016; 35: 1457-1463.
- 62) Konstadt S, Reich D, Rafferty T. Validation of a test of competence in transesophageal echocardiography. *J Cardiothoracic Vasc Anesth* 1996; 10(3): 311-313.
- 63) Sheehan F, Otto C, Freeman R. Echocardiography simulator with novel training and competency testing tools. *Studies in health tech and informatics* 2013; 184: 397-403.

- 64) Damp J, Anthony R, Davidson M et al. Effects of Transesophageal Echocardiography Simulator Training on Learning and Performance in Cardiovascular Medicine Fellows. *J Am Soc Echocardiography* 2013; 26(12): 1450-6.
- 65) Bick J, DeMaria S, Kennedy J et al. Comparison of Expert and Novice Performance of a Simulated Transesophageal Echocardiography Examination. *Sim Healthcare* 2013; 8: 329-334.
- 66) Beraud AS, Rizk NW, Pearl RG et al. Focused transthoracic echocardiography during critical care medicine training: curriculum implementation and evaluation of proficiency*. *J. Crit Care Med.* 2013 Aug;41(8):e179-81.
- 67) Sohmer B, Hudson C, Hudson J et al. Transesophageal echocardiography simulation is an effective tool in teaching psychomotor skills to novice echocardiographers. *Can J Anesth* 2014; 61 (3): 235-41.
- 68) Jelacic S, Bowdle A, Togashi K et al. The Use of TEE Simulation in Teaching Basic Echocardiography Skills to Senior Anesthesiology Residents. *J Cardiothorac and Vasc Anesth* 2013; 27(4): 670-5.
- 69) Edrich T, Seethala R, Olenchock B et al. Providing Initial Transthoracic Echocardiography Training for Anesthesiologists: Simulator Training Is Not Inferior to Live Training. *J Cardiothorac and Vasc Anes* 2014; 28(1): 49-53.
- 70) Cawthorn T, Nickel C, O'Reilly M et al. Development and evaluation of methodologies for teaching focused cardiac ultrasound skills to medical students. *J Am Soc Echocardiography* 2014 27(3): 302-9.
- 71) Gulbrand Nielsen D, Jensen S, O'Neill L. Clinical assessment of transthoracic echocardiography skills: a generalizability study. *BMC Medical Education* 2015; 15(9): 1-7.
- 72) Arntfield R, Pace J, Mcleod S et al. Focused transesophageal echocardiography for emergency physicians – description and results from simulation training of a structured four-view examination. *Critical Ultrasound Journal* 2015; 7(10): 1-7.
- 73) Ho A, Critchley MH, Lester AH et al. Introducing final-year medical students to pocket-sized ultrasound imaging: Teaching transthoracic echocardiography on a 2-week anesthesia rotation *Teaching and Learning in Medicine* 2015; 27(3): 307-313.
- 74) Markowitz J, Hwang J, Moore C. Development and Validation of a Web-Based Assessment Tool for the Extended Focused Assessment With Sonography in Trauma Examination. *J Ultrasound Med* 2011; 30: 371-5
- 75) Hofer M, Kamper L, Sadlo M et al. Evaluation of an OSCE assessment tool for abdominal ultrasound courses. *Ultraschall Med.* 2011 Apr;32(2):184-90

- 76) Thoires K and Coffee J. Developing the clinical psychomotor skills of msk sonography using a multimedia DVD. *Australasian J Educ Tech* 2012, 28 (4) 703-718.
- 77) Madsen ME, Konge L, Nørgaard LN et al. Assessment of performance measures and learning curves for use of a virtual reality ultrasound simulator in transvaginal ultrasound examination. *Ultrasound Obstet Gynecol*. 2014 Dec;44(6):693-9.
- 78) Jaffer U, Pandey VA, Aslam M et al. Validation of a novel duplex ultrasound objective structured assessment of technical skills (DUOSATS) for arterial stenosis detection. *Heart Lung and Vessels* 2014; 6(2): 92-104.
- 79) Tolsgaard MG, Ringsted C, Dreisler E et al. Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. *Ultrasound Obstet Gynecol* 2014; 43: 437-443.
- 80) Jaffer U, Normahani P, Lackenby K, et al. Validation of a Novel Venous Duplex Ultrasound Objective Structured Assessment of Technical Skills for the Assessment of Venous Reflux. *J Surg Educ* 2015; 72(4): 754-60.
- 81) Todsén T, Tolsgaard M, Olsen B et al. Reliable and Valid Assessment of Point-of-Care Ultrasonography. *Annals of Surgery* 2015; 261(2): 309-315.
- 82) Ziesmann M, Park J, Unger B et al. Validation of the quality of ultrasound imaging and competence (QUICK) score as an objective assessment tool for the FAST examination. *J Trauma Acute Care Surg* 2015; 78(5): 1008-1013.
- 83) Ziesmann M, Park J, Unger B et al. Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. *J Trauma Acute Care Surg* 2015; 79(4): 631-7.
- 84) Chaudery M, Clark J, Dafydd D, et al. The Face, Content, and Construct Validity Assessment of a Focused Assessment in Sonography for Trauma Simulator. *J Surg Educ* 2015; 72(4): 1032-8.
- 85) Patrawalla P, Wise W, Eisen LW, et al. Development and Validation of an Assessment Tool for Competency in Critical Care Ultrasound. *J Grad Med Education* 2015; 567-73.
- 86) Schmidt JN, Kendall J, Smalley C. Competency Assessment in Senior Emergency Medicine Residents for Core Ultrasound Skills. *West J Emerg Med*. 2015 Nov;16(6):923-6.
- 87) Dyre L, Nørgaard LN, Tabor A, et al. Collecting Validity Evidence for the Assessment of Mastery Learning in Simulation-Based Ultrasound Training. *Ultraschall Med*. 2016 Aug;37(4):386-92.

- 88) Amini R, Stolz LA, Javedani PP, et al. Point-of-care echocardiography in simulation-based education and assessment. *Adv Med Educ Pract* 2016 May 31;7:325-8.
- 89) Black H, Sheppard G, Metcalfe B et al. Expert Facilitated Development of an Objective Assessment Tool for Point-of Care Ultrasound Performance in Undergraduate Medical Education. *Cureus*. 2016 Jun 10;8(6):e636.
- 90) Tolsgaard MG, Todsen T, Sorensen J, et al International Multispecialty Consensus on How to Evaluate Ultrasound Competence: A Delphi Consensus Survey. *PLoS ONE* 2013; 8(2): e57687.
- 91) Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. *Adv Health Sci Educ Theory Pract*. 2014. Epub ahead of print 2014 Sep 27.
- 92) Setna Z, Jha V, Boursicot KAM et al. Evaluating the utility of workplace-based assessment tools for specialty training. *Best Pract Res Clin Obstet Gynaecol* 2010, 24(6):767–82.
- 93) Larsen CR, Grantcharov T, Aggarwal R et al. Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surg Endosc* 2006, 20(9):1460–6.
- 94) Siddiqui NY, Stepp KJ, Lasch SJ et. al. Objective structure assessment of technical skills for repair of fourth-degree perineal lacerations. *Am J Obstet Gynecol* 2008, 199(6):676.e1–676.e6.
- 95) Gofton W, Dudek N, Wood T et al. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. *Acad Med* 2012; 87: 1401-7.
- 96) Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206–214.
- 97) Fernando N, Cleland J, McKenzie H, Cassar K. Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. *Med Educ*. 2008;42:89 – 95.
- 98) Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the mini clinical evaluation exercise. *J Gen Intern Med*. 2004;19:558 –561.
- 99) Watt D. On becoming a qualitative researcher: the value of reflexivity. *Qual Report* 2007; 12(1): 82-101.
- 100) Hsieh H, Shannon S. Three approaches to qualitative content analysis. *Qual Health Res* 2005; 15(9): 1277-1288.

- 101) Harrison C, Konigs K, Schuwirth L et al. Barriers to the uptake and use of feedback in the context of summative assessment. *Adv Health Sci Educ* 2015; 20: 229-45.
- 102) Ende J. Feedback in clinical medical education. *JAMA* 1983; 250; 777-81.
- 103) McGaghie WC, Issenberg SB, Petrusa ER et al. A critical review of simulation- based medical education research: 2003–2009. *Medical Education*. 2010; 44(1), 50–63.
- 104) Humphrey-Murto S, Pugh D, Touchie C et al. Feedback in the OSCE: what do residents remember? *Teach Learn Med* 2016, 28(1): 52-60.
- 105) Hatala R, Cook D, Zendejas B et al. Feedback for simulation-based procedural skills training: a meta-analysis and critical narrative synthesis. *Adv Health Sci Educ Theory Pract*. 2014;19(2):251-72.
- 106) Schmidt R, Lee T. *Motor control and learning: A behavioral emphasis* (5th ed.).2011; Champaign: Human Kinetics.
- 107) Van de Ridder JM, Berk F, Stokking K et al. Feedback providers' credibility impacts students' satisfaction with feedback and delayed performance. *Medical Teacher*. 2005; 37(8): 767-774.
- 108) Dieckmann P, Gaba D, Rall M. Deepening the Theoretical Foundations of Patient Simulation as Social Practice. *Sim Healthcare* 2007; 2:183–193.
- 109) Tavares W, Brydges R, Myre P et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. *Adv Health Sci Educ Theory Practice* 2017 [Epub ahead of print].
- 110) Bordage G. Conceptual frameworks to illuminate and magnify. *Medical Education*. 2009; (43)4: 312-319.

APPENDIX A – CoCATS 4 Task Force 5 Core Competency Components and Curricular Milestones for training in echocardiography

TABLE 1 Core Competency Components and Curricular Milestones for Training in Echocardiography

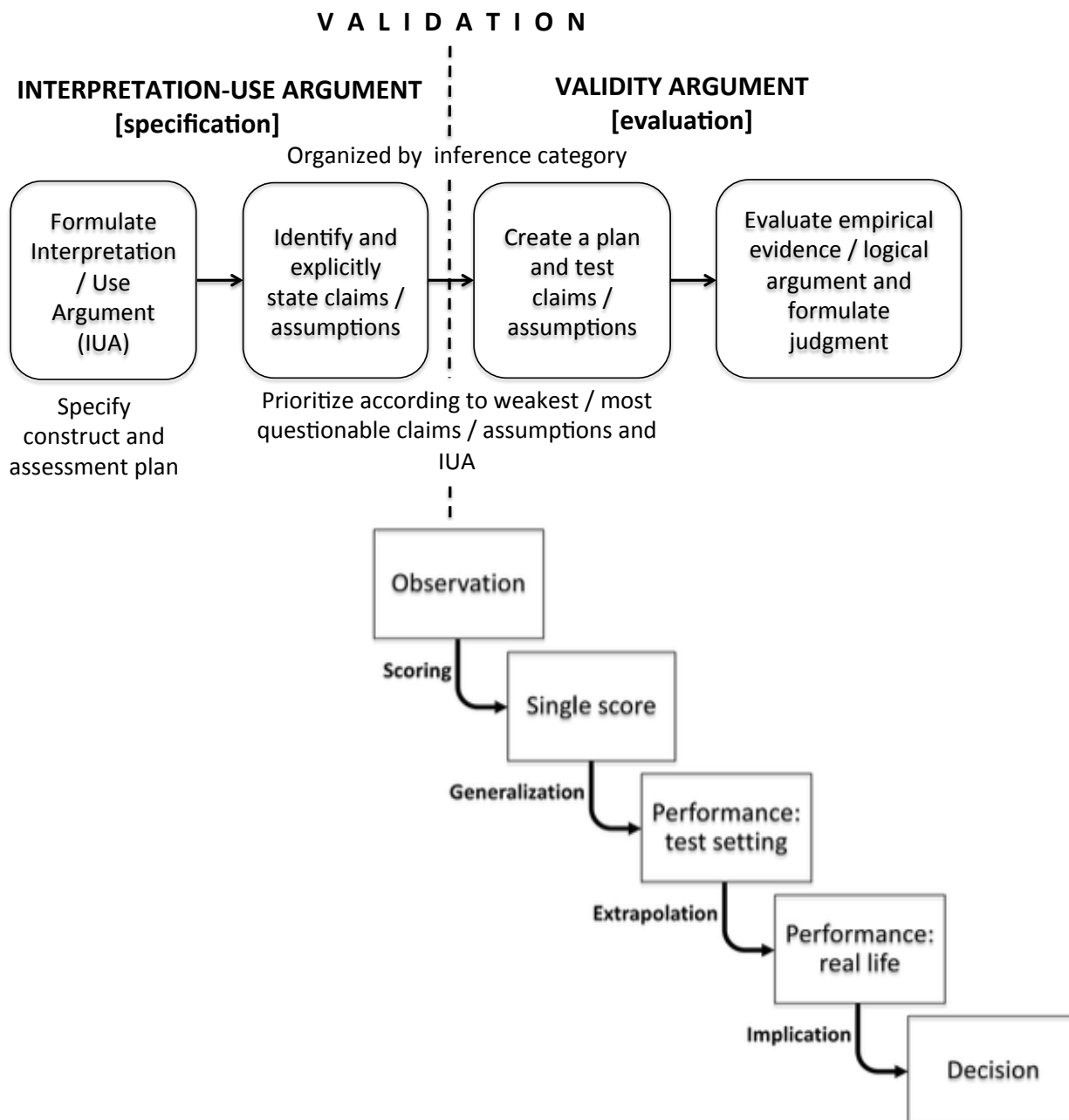
Competency Component		Milestones (Months)			
MEDICAL KNOWLEDGE		12	24	36	Add
1	Know the physical principles of ultrasound and the instrumentation used to obtain images.	I			
2	Know the appropriate indications, including the appropriate use criteria, for: M-mode, 2-dimensional, and 3-dimensional transthoracic echocardiography; Doppler echocardiography and color-flow imaging; transesophageal echocardiography; tissue Doppler and strain imaging; and contrast echocardiography.		I		
3	Know the limitations and potential artifacts of the echocardiographic examination.	I			
4	Know the standard views included in a comprehensive transthoracic echocardiogram.	I			
5	Know the standard views included in a comprehensive transesophageal echocardiogram.		I		
6	Know the techniques to quantify cardiac chamber sizes and evaluate left and right ventricular systolic and diastolic function and hemodynamics.			II	
7	Know the characteristic findings of cardiomyopathies.		I		
8	Know the use of echocardiographic and Doppler data to evaluate native and prosthetic valve function and diseases.			II	
9	Know the echocardiographic and Doppler findings of cardiac ischemia and infarction, and the complications of myocardial infarction.		I		
10	Know the echocardiographic findings of pericardial disease, pericardial effusion, and pericardial constriction.		II		
11	Know the characteristic findings of basic adult congenital heart disease.			II	
12	Know the findings of complex/postoperative adult congenital heart disease.			III*†	III*
13	Know the techniques to evaluate cardiac masses and suspected endocarditis.		II		
14	Know the techniques to evaluate diseases of the aorta.		II		
15	Know the techniques to assess pulmonary artery pressure and diseases of the right heart.		II		
16	Know the use and characteristic findings in the evaluation of patients with systemic diseases involving the heart.		II		
17	Know the indications for, and the echocardiographic findings in, patients with known or suspected cardioembolic events.		II		
18	Know key aspects of contrast echocardiography including interpretation, administration techniques, and safety information.			II	
19	Understand the principles and applications of 3-dimensional echocardiography.		II		
20	Recognize and treat the potential complications of stress, contrast, and transesophageal echocardiography.		II		
EVALUATION TOOLS: conference presentation, direct observation, and in-training examination.					
PATIENT CARE AND PROCEDURAL SKILLS		12	24	36	Add
1	Skill to perform and interpret a basic transthoracic echocardiographic examination.		I		
2	Skill to perform and interpret a comprehensive transthoracic echocardiographic examination.			II	
3	Skill to perform and interpret a comprehensive transesophageal echocardiographic examination.			II	
4	Skill to recognize pathophysiology, quantify severity of disease, identify associated findings, and recognize artifacts in echocardiography.			II	
5	Skill to integrate echocardiographic findings with clinical and other testing results in the evaluation and management of patients.		I		
6	Skill to interpret stress echocardiography.			II	
7	Skill to incorporate stress hemodynamic information in the management of complex valve disease or hypertrophic cardiomyopathy.			II	
8	Skill to utilize echocardiographic techniques during cardiac interventions, including intraoperative transesophageal echocardiography.			III†	III
9	Skill to perform and interpret basic 3-dimensional echocardiography.			II	
10	Skill to utilize advanced 3-dimensional echocardiography during guidance of procedures and/or surgery.			III†	III
11	Skill to perform and interpret contrast echocardiographic studies.			II	
EVALUATION TOOLS: direct observation, logbook, and simulation.					

TABLE 1 Core Competency Components, continued

Competency Component		Milestones (Months)			
SYSTEMS-BASED PRACTICE		12	24	36	Add
1	Work effectively and efficiently with the echocardiography laboratory staff.	I			
2	Incorporate risk/benefit, safety, and cost considerations in the use of ultrasound techniques.			I	
3	Participate in echocardiographic quality monitoring and initiatives.			II	
EVALUATION TOOLS: direct observation and multisource evaluation.					
PRACTICE-BASED LEARNING AND IMPROVEMENT		12	24	36	Add
1	Identify knowledge and performance gaps and engage in opportunities to achieve focused education and performance improvement.		I		
EVALUATION TOOLS: conference presentation and direct observation.					
PROFESSIONALISM		12	24	36	Add
1	Know and promote adherence to guidelines and appropriate use criteria.		I		
2	Interact respectfully with patients, families, and all members of the healthcare team, including ancillary and support staff.	I			
EVALUATION TOOLS: conference presentation, direct observation, multisource evaluation, and reflection and self-assessment.					
INTERPERSONAL AND COMMUNICATION SKILLS		12	24	36	Add
1	Communicate with and educate patients and families across a broad range of cultural, ethnic, and socioeconomic backgrounds.		II		
2	Communicate testing results to physicians and patients in an effective and timely manner.		II		
3	Communicate detailed information on cardiac anatomy for surgical planning or guidance of interventional procedures.			II	
EVALUATION TOOLS: direct observation and multisource evaluation.					

Ryan et al 2015 [24]. Reprinted from *Journal of the American College of Cardiology*, 65 (17) Ryan T, Berlacher K, Lindner J et al. COCATS 4 Task Force 5: Training in Echocardiography, pp1786-1799m 2015 with permission from Elsevier.

APPENDIX B Conceptual Framework: Kane's Argument-Based Approach to Validity



Combined diagrams from Tavares et al. 2017 [111] Cook et al. 2016 [41]. Reprinted by permission from Springer Link: Advances in Health Science Education, Applying Kane's validity framework to a simulation based assessment of clinical competence, Tavares et al, 2017.

APPENDIX C - Kane's Argument-Based Approach to Validity Conceptual Framework: Evidence

Inference	Operational sources of evidence to support the validity argument	
	Quantitative	Qualitative
Scoring <i>Is the rule applied as specified?</i> <i>Are raters appropriately selected to assess trainee performance in the domains?</i> <i>Are raters properly trained to provide consistent and accurate ratings?</i>	<ul style="list-style-type: none"> Item and response option performance (item difficulty, point biserial, response option analyses) Observation format (e.g., empiric comparison of different formats, such as live vs video-based, or blinded vs unblinded scoring) Standardization, equating Scoring rubric/criteria (e.g., empiric comparison of different procedures, think-aloud study) Rater selection and training; rater accuracy and reliability Data security, quality control 	<ul style="list-style-type: none"> Observations actually conducted The richness, accuracy, authenticity, and fairness of qualitative data (e.g. individual narratives, other documents)
Generalization <i>Is the sample of observations appropriately representative of the universe of possible observations? Is the sample size large enough to control for random error or sampling error?</i> <i>What are the results of reliability or generalizability analyses?</i>	<ul style="list-style-type: none"> Reliability / generalizability (items, raters, tasks, occasions) Item-response theory 	<ul style="list-style-type: none"> Sampling and triangulation; the variety of perspectives reflected in data being analyzed (different observers, performance domains, time points, data types) Defensibility, reflexivity, transparency, and responsiveness of the interpretive process Thematic saturation and coherence of final interpretations Consistency and reflexivity of interpretations formed by different interpreters
Extrapolation <i>How do observed ratings correlate with real-world outcomes of interest?</i> <i>How do observed ratings correlate with other methods of assessing similar constructs?</i> <i>Are there artificial aspects of testing conditions that affect</i>	<ul style="list-style-type: none"> Needs analysis to define scope/objectives Process-construct match (e.g., think-aloud study) Relevance and authenticity (e.g., ratings by experts) Correlation with another measure having an expected relationship (criterion-referenced or convergent; concurrent or predictive) 	<ul style="list-style-type: none"> Agreement of relevant stakeholders (e.g., observers, learners, program directors) with final interpretation (member check) Agreement of stakeholders that interpretations will apply to new contexts in training or practice (transferability) Relationship between qualitative interpretations and

<i>trainee performance and ratings?</i>	<ul style="list-style-type: none"> • Discrimination (known groups comparison) • Responsiveness (sensitivity to change following intervention) • Construct profile (e.g., factor analysis, multi-trait multimethod matrix) • Differential item functioning • The relevance of data sources to performance 	other measures of similar traits (e.g., quantitative data, independent decisions about remediation or honors)
Implications <i>What is the impact on assessment on the learner, the program, and society?</i>	<ul style="list-style-type: none"> • Pass/fail standard (e.g. receiver-operating characteristics curve) • Effectiveness of actions based on assessment results • Intended or unintended consequences of testing (long-term follow-up; qualitative studies; consider impact on learners, raters, and others) 	<ul style="list-style-type: none"> • Agreement of other experts with final judgment and decision • Effectiveness of actions based on assessment results • Intended or unintended consequences of testing (consider impact on learners, observers, interpreters, and others)

Adapted from: Cook et al. Medical Education 2015 [41], Hawkins et al. Acad Medicine 2010 [55], Hatala et al. Adv Health Sci Educ 2015 [58]

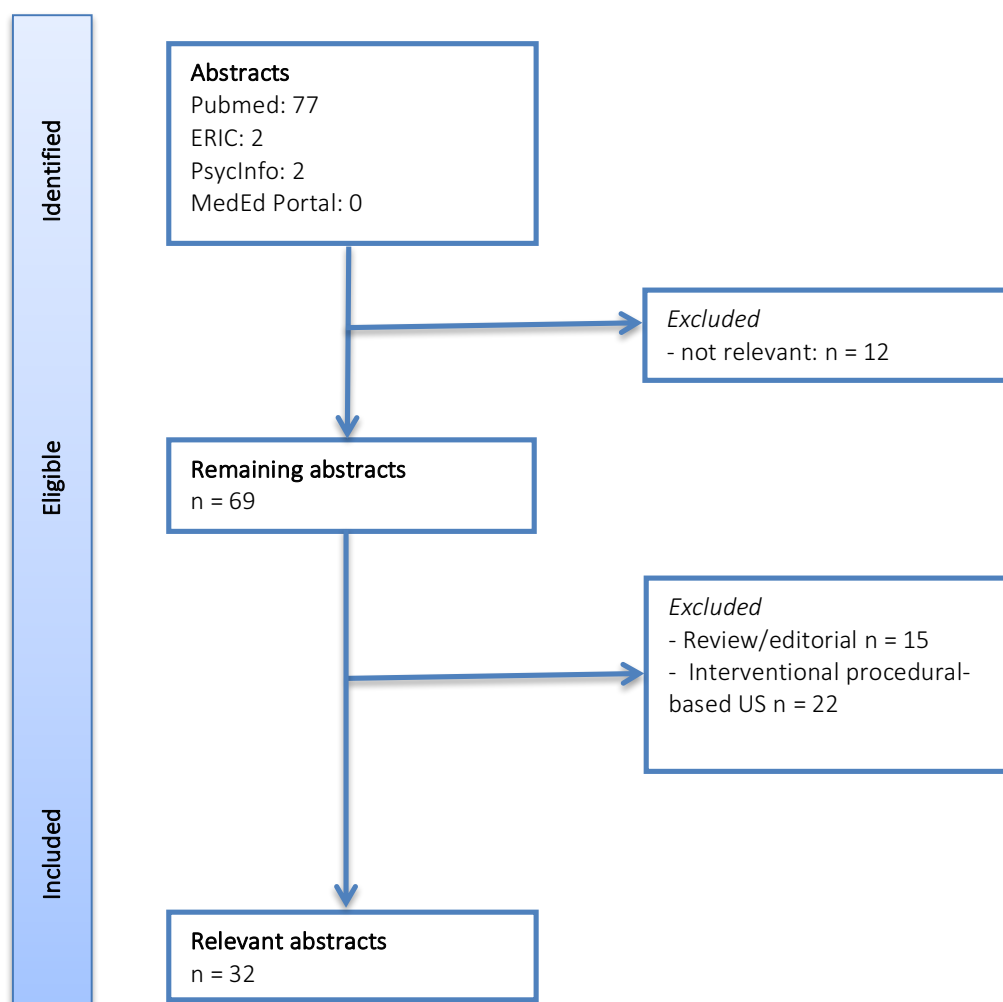
APPENDIX D - Operationalizing Kane's Argument-Based Approach to Validity

1. Define the construct and proposed interpretation of the assessment data.
2. Make explicit the intended decision(s).
3. Define the interpretation/use argument of assessment data, and prioritize needed validity evidence.
4. Identify candidate instruments and/or create/adapt a new instrument.
5. Appraise existing evidence and collect new evidence as needed.
6. Keep track of practical issues including cost.
7. Formulate/synthesize the validity argument in relation to the interpretation/use argument.
8. Make a judgment: does the evidence support the intended use?

Cook and Hatala. Advances in Simulation 2016 [45]

APPENDIX E Expanded Literature Review:

Key words: (echo, echocardiography, transthoracic, transesophageal, ultrasound) AND (competence, competency) AND (assessment, tool, validity, validation)



APPENDIX F Validity evidence in the literature for assessments of competency using ultrasound

Study	Assessment description	Proposed interpretation/use	Inf	Evidence for Inferences (S = Scoring, G = Generalization, E = Extrapolation, I = Implications)	Favorable/unfavorable; weak/strong
ECHO/FOCUSED CARDIAC US					
Konstadt et al. J Cardiothorac and Vasc Anesth 1996 [63]	Anesthesia 11 residents and 14 faculty assessed. 34 video loops with MCQ. Residents repeated MCQ after 1yr of clinical training.	Standardized MCQ test performance measures intraoperative echo diagnostic competence	S	- Scoring rubric? – written by experienced anesthesia echo experts, with unambiguous videos of normal and pathologic images, used teaching file *unclear if blueprint - Observation format? no observation: MCQ numerical response in K-type question; * single best answer could discriminate more - Rater selection/training? n/a as MCQ - Rater consistency/accuracy? Theoretically high as MCQ - Qualitative narratives? Not done	Favourable, weak
			G	- Representative sampling? – unclear how questions selected, only tested transverse not biplane/multiplane/epicardial/spectral doppler * many aspect raised as study limitations - error? N/a given MCQ - Reliability/generalizability analyses? n/a, * study limitation	Favourable, weak
			E	- Real-world correlation? Residents scores improves after 1 yr of training. - Scores discriminate groups? Residents scored lower than faculty before training. After training, no difference between groups.	Favourable, moderate
			I	- impact on learner, program, society? Previously used as self-assessment for residents and quality assurance but not discussed - standard setting? n/a	Not clear
Neelankanvil et al. Anesthesia Analgesia 2012 [37]	Anesthesia 61 residents assessed with written 20 MCQ pre-test, 20 MCQ post-test and observed post-test TTE (5 views in parasternal and apical measuring 5-item quality, binary structures, time) on	Observed TTE testing measures procedural competence in a simulation vs traditional teaching educational intervention ** educational efficacy study, not	S	- Scoring rubric? – per FATE protocol, inc quality of view (0-5) and assessment of structures Y/N * subjective - Observation format? direct obs and offline - Rater selection/training? 2 expert anesthetists, * training n/a - Rater consistency/accuracy? Raters blinded to grp assignment and (offline) to identity. IRR 0.83 for first session, 0.87 for second	Favourable, moderate strength
			G	- Representative sampling? 2 imaging window of highest yield to anesthetist in ICU/postop by FATE (PS and Ap) w 5 views (Lax, RV infl, PS PM, A4C, A2C). Same level difficulty * limitation - error? Controlled by replicating post-test on one volunteer 61 times after first session, and a second volunteer 21 times after second session - Reliability/generalizability analyses? Not done	Favourable, moderate strength

	<p>patient. Educational intervention was lecture + video vs. sim.</p> <p>After 3w, subset n= 21 assessed with MCQ post-test, observed TTE on patient. Educational intervention was hands-on patient scanning vs sim.</p>	assessment validity study	E	<p>- Real-world correlation? Compared to control group, Sim group performed better at image quality and anatomy identification score, took less time per view, higher % correctly obtained view.</p> <p>- Scores discriminate groups? training level inconsistent data (CA yr1 and CA yr3 better image quality sim than control but CA yr2 n/s; all better at anatomy identification. Did not compare results from session 1 and 2 systematically)</p>	Favourable, moderate strength
			I	<p>- impact on learner, program, society? Short-term educational approach but long-term retention and clinical application not known</p> <p>- standard setting? Saw that after sim training could acquire images and identify structures >90% which could help standards</p>	Favourable, weak
Sheehan et al. Studies in health technology and informatics 2013 [64]	<p><u>Medicine/Cardio</u></p> <p>18 medicine residents/novices and 6 echo experts trained with didactic + 7 sim cases, then assessed with 9 MCQ and observed TTE on sim (5-6, plane error automated)</p>	A quantitative objective assessment measures psychomotor and cardiac ultrasound skills	S	<p>- Scoring rubric? Not discussed</p> <p>- Observation format? on simulator, % plane error to 3D ideal</p> <p>- Rater selection/training? Automatic rater - sim</p> <p>- Rater consistency/accuracy? Automatic rater – sim thus high reproducibility of measurements</p>	Favourable, very weak
			G	<p>- Representative sampling? Rationale for 6 views selected not provided</p> <p>- Reliability/generalizability analyses? n/a</p>	Favourable, very weak
			E	<p>- Real-world correlation? Negative correlation of mean angle error and cognitive skill ($r=-0.47$)</p> <p>- Scores discriminate groups? novices vs experts: experts had lower error on image acquisition of 2 of 6 views but novices did as well with other 4 views. Image interpretation experts scored higher.</p>	Inconsistent, weak
			I	<p>- impact on learner, program, society? Intent to provide validated tool for competency testing rather than like an OSATs for formative training but propose for full spectrum of training and applications (benchmark performance, enable training to proficiency, formative assessment, acceleration of skill acquisition and skill retention) * not supported by data above</p>	Favourable, weak
Nielsen et al. BMC Medical Education 2013 [61]	<p><u>Medicine/Cardio</u></p> <p>15 interns, 15 cardio residents, 15 experts assessed by objective assessment of</p>	An objective assessment of TTE technical skills measures TTE procedural competence.	S	<p>- Scoring rubric? ? framework based on literature, national and international guidelines, author expert consensus, pilot</p> <p>- Observation format? Video review with GRS and CL</p> <p>- Rater selection/training? Author defined criteria for rating and gave thorough intro, observing rating of first 15 to ensure rating consensus; one expert rated 45 exams * limitation</p>	Favourable, strong

	technical echo skills on patient (CL+ 5-item GRS). Pilot of 9 MDs scanning 3 pts.			- Rater consistency/accuracy? More reliable as images produced assessed rather than performance so more objective. Expert rated in random order, blinded to identity, de-identified exams. Re-rated 10 of same 4 wks later. 2 nd expert graded those 10 as well; Intra-rater 0.67 GRS, 0.99 checklist. Inter-rater 0.61 GRS, 0.95 checklist. Collapsed checklist mean score /5 not /440 like GRS has intra-rater 0.83, inter-rater 0.66.	
			G	- Representative sampling? Exam not sampling – comprehensive new instrument designed for this study, with all items required in standard Danish TTE scan 440 points *feasibility limitation. In pilot, assessed 3 cases with escalating difficulty. - Reliability/generalizability analyses? See below, Gulbrand Nielsen et al. 2015 - Measurement error? 45 exams on same patient *limitation to feasibility but reduces error	Favourable, strong
			E	- Real-world correlation? Mean performance time strong negative correlation w expertise: interns longest, residents intermediate, experts shortest. In pilot, for 2 most difficult cases novices and residents take same time and experts shortest. - Scores discriminate groups? Significant correlation b/w expertise level and GRS (r=0.76) and total CL (r=0.74). In pilot, residents and consultants n/s difference on GRS and on CL, though correlation still found w expertise and GRS (r=0.70) *limitation ?lack of discrimination from power vs selection bias - measurement error/other – significant correlation b/w GRS and CL (0.88) suggesting measuring same trait though GRS should reward efficacy and CL step-by-step approach. Novices assisted with machine could have attenuated differences *limitation	Generally favorable with some inconsistencies, strong
			I	- impact on learner, program, society? Proposed future directions but no evidence provided - standard setting? CL highly reliable (ICC>0.8 acceptable for high stakes)	Favourable, weak
Damp et al. JASE 2013 [65]	<u>Cardio</u> 19 residents assessed with observed TEE on patient (3-item GRS for 37 views/ structures) + self-	Observed TEE testing measures technical procedural competence in a simulation training vs. standard	S	- Scoring rubric? ASE complete TEE as framework - Observation format? observed - Rater selection/training? 2 expert faculty, training not discussed - Rater consistency/accuracy? Same 2 experts rated for consistency but nonblinded and no inter/intra-rater discussed * - Qualitative? Self-assessment questionnaire has observer insights on confidence and utility of training	Favourable, weak but some unique aspects

	assessment questionnaire (5-item scale for 9 Q) crossover design with either regular sim practice for month 1 or month 2 of ttl 2 month rotation. 8 residents in historical control assessed with TEE assessment on patient.	educational intervention? ** educational efficacy study, not assessment validity study	G	- Representative sampling? No sampling – complete TEE exam per ASE w all aspects doppler, angulation. Difficulty: observed TEEs on clinical patients with potential for differing difficulty *limitation and source of potential error - Reliability/generalizability analyses? n/a	Favourable, moderate
			E	- Real-world correlation? Compared to control, sim group had higher assessment scores, # views achieved without assistance, lower # of views not achieved. Inconsistency as for ME and Ao no difference bw control and sim though difference in Deep esop and deep TG, and n/s trend to shorter perform time *limitation/true differences as easier views? Higher scores after month 1 for sim month 1 group, # views achieved without assistance and fewer # of views with instruction suggesting sim incrementally beneficial. Skill set maintained w smaller change in ttl score by month 2 c/w control. Not able to correlate time spent on sim w scores due to power *limitn - Scores discriminate groups? n/a as homogeneous group	Favourable but some inconsistencies, Strong
			I	- impact on learner, program, society? Self-reported time on simulator can help anticipate program needs. Discuss feedback at the time might help. Discuss minimum time on sim to detect score improvement. Note: bw sim groups no difference in perceptions of training, but earlier sim group had more comfort w TEE, confidence to perform TEE, comfort w abN result interpretation did not compare w control *limitation. - standard setting? n/a	Favourable with some inconsistencies, weak
Bick et al. Simul Healthcare 2013 [66]	Anesthesia 15 novice residents and 11 experts assessed by BTEET on sim (2 parameters for each of 10 views: acquisition time and interpretation)	A standardized Basic TEE Evaluation Tool assesses procedural competence	S	- Scoring rubric? Standardized basic transesophageal echo evaluation tool developed by experts at Vanderbilt and Mt Sinai - Observation format? digital AV capture, voice masking, offline - Rater selection/training? 3 TEE-certified anesthesiologists but image quality “I know it when I see it” inconsistent results so couldn't use *limitation - Rater consistency/accuracy? Blinded, off-site to acquisition, Voice masking	Favourable, moderate
			G	- Representative sampling? 10 standard TEE planes selected, aspects image and verbal identification for structures/doppler if relevant - Reliability/generalizability analyses? n/a	STRONGEST
			E	- Real-world correlation? Image view time significant quicker for experts c/w novices for 9 of 10 views (not mid esop AA)	Inconsistencies, weak

				- Scores discriminate groups? IV score results not provided and n/s difference experts and novices for structural anatomy score for 5 of 10 views *limitation	
			I	- impact on learner, program, society? Some aspects help discriminate but based on this can't assess multiple points from novice to expert	Favourable, weak
Beraud et al. Crit Care Med 2013 [67]	<u>Critical Care</u> 18 residents assessed by quality and accuracy of images in 3 pts, standardized sim exam (5 pathologic scenarios measuring time to acquire 5 views correctly), and a MCQ, after a 1 year F-TTE curriculum (lectures, scanning instruction). Pilot of 9 novices and 5 experts for sim assessment.	Observed focused TTE testing measures proficiency after an educational intervention ** educational efficacy study, not assessment validity study	S	- Scoring rubric? Mapped to curricular objectives and FTTE consensus statement training requirements. Metric = time. Observation format? direct observation. - Rater selection/training? Single instructor recorded all - Rater consistency/accuracy? No comment by rater on performance, score focused on diagnosis time, highly reproducible as computerized	Favourable, strong but limited
			G	- Representative sampling? Scenarios represented spectrum of difficulty but limited metric assessed (time to make diagnosis) - Reliability/generalizability analyses? n/a	Favourable, moderate
			E	- Real-world correlation? Unable to assess correlation bw duration of training and acquisition of proficiency - Scores discriminate groups? Pilot testing: diagnosis time score for all 5 and individual clinical scenarios increases from expert to fellow to novice	Favourable, moderate
			I	- impact on learner, program, society? Unclear as limited metric - standard setting? Suggest "a proficiency evaluation system may be an appropriate tool for determining when fellows have been adequately prepared for use of F-TTE in independent practice" but not supported	Inconsistent, weak
Sohmer et al. Can J Anesth 2013 [68]	<u>Anesthesia</u> 33 TEE-naive anesthesiologists assessed with observed pre and post-test (4-item scale for 10 views) and pre and post-test MCQ. Educational intervention: instructor guided sim vs self-directed sim.	Observed focused TEE testing measures psychomotor procedural competence in an educational intervention ** educational efficacy study, not assessment validity study	S	- Scoring rubric? Per TEE NBE guidelines, developed through iterative mod Delphi w 4 TEE experts, 3 rounds, w image quality and able to use/not use for diagnostic interpretation - Observation format? video - Rater selection/training? 2 NBE TEE experts. Training not discussed * limitn - Rater consistency/accuracy? Raters blinded to each other, instructional modality, phase of testing. High reliability w ICC 0.98	Favourable, strong
			G	- Representative sampling? 10 of 20 images required for TEE previously described as most important images to acquire. Normal mode only. - Reliability/generalizability analyses? n/a	Favourable, moderate
			E	- Real-world correlation? Both instructional modalities improved scores with no difference between groups. - Scores discriminate groups? n/a	Favourable, weak

			I	- impact on learner, program, society? Educational intervention posited as an adjunct, could reduce learning bottleneck, lack of inferiority of self-directed suggest possible directions but evidence does not relate to assessment tool	Favourable, weak
Jelacic et al. J Cardiothoracic and Vasc Anesth 2013 [69]	Anesthesia 37 residents assessed by 25-item MCQ pre-test and post-test 26d later. Educational intervention was simlab TEE tutorial. 9 faculty underwent MCQ testing.	Standardized MCQ test performance measures intraoperative echo cognitive competence after educational intervention ** educational efficacy study, not assessment validity study	S	- Scoring rubric? Limited to cognitive skills safety, probe manipulation, anatomy, applications, basic pathology. Not clear how defined but administered to 7 faculty and reviewed with 2 senior subject matter experts. - Observation format? MCQ - Rater selection/training? n/a - Rater consistency/accuracy? n/a - qualitative – narrative responses to training course provided, but not related to assessment testing	Favourable, weak
			G	- Representative sampling? Unclear how subject matter selected - Reliability/generalizability analyses? n/a	Weak
			E	- Real-world correlation? No significant improvement in global score b/w pre and post-test, though improved knowledge of anatomy by residents - Scores discriminate groups? Faculty score pre and post-test both significantly higher than residents.	Weak
			I	- impact on learner, program, society? Program valued by residents but evidence does not relate to assessment tool	Favourable, weak
Edrich et al. J Cardiothor Vasc Anest 2014 [70]	Anesthesia 46 TTE-novices assessed by 40MCQ written and practical (time to acquire 5 standard views, 4-item quality points per view) pre-and post-tests. Educational intervention was sim vs live volunteer.	Observed TTE test performance measures intraoperative echo psychomotor competence in a simulation vs standard clinical educational intervention ** educational efficacy study, not assessment validity study	S	- Scoring rubric? Per periop Cric Care US/ACC/ACEP trainin expert authors given quality points - Observation format? video clips - Rater selection/training? TTE experts, training n/a * limitn - Rater consistency/accuracy? Test administrator acquired, de-identified clips, clips score separately in random order, single volunteer scanned 46 times. IRR alpha 0.85 “near-perfect”	Favourable, moderate
			G	- Representative sampling? Unclear how 5 views selected of possible that could have been, unclear how determined quality points to assess. Single volunteer improved reproducibility but single level of difficulty. - Reliability/generalizability analyses? n/a	Favourable, weak
			E	- Real-world correlation? Image-acquisition scores improved in both groups from pre-to post-training. Attempt to compare the groups found non-inferior sim training vs live - Scores discriminate groups? n/a as homogeneous group	Favourable, weak

			I	- impact on learner, program, society? Educational intervention proposed to address “scalable and efficient methods to train anesthetists” and “challenges of scheduling for trainees and volunteers” however this evidence is not assessment related	Better
Cawthorn et al. J Am Soc Echo [71]	<u>Undergraduate Medical Education</u> Phase 1 n=12, phase 2 n=45, students assessed by hand held ultrasound on pt (Phase 1, 4-item image quality for 7 view; Phase 2, 9-items scan quality and accuracy for 8 views). Educational intervention was (Phase 1) didactic course + practical HHU and (Phase 2) didactic course + electronic modules + practical HHU	Observed structured echo assessment measures hand-held cardiac ultrasound image acquisition proficiency. ** educational efficacy study, not assessment validity study	S	- Scoring rubric? Expert defined, subjective image quality and diagnostic accuracy not clearly articulated - Observation format? video review - Rater selection/training? Phase 1 – single ASE level III echo rater; phase 2 – 3 ASE level III raters - Rater consistency/accuracy? Raters blinded to de-identified studies, single rater in phase 1; phase 2 ICC 0.73-0.92	Favourable, moderate
			G	- Representative sampling? Not clear how parameters selected * normal patient thus not assessing difficulty - Reliability/generalizability analyses? n/a	Favourable, weak
			E	- Real-world correlation? Testing not done pre/post thus n/a - Scores discriminate groups? Homogeneous group thus n/a	n/a
			I	- impact on learner, program, society? Assessment as part of formative process - standard setting?	Favourable, weak
Gulbrand Nielsen, MBC Med Educ 2015 [72]	<u>Cardio</u> 3 novices, 3 cardiology residents, 3 faculty assessed by objective assessment of technical echo skills on 3 pts (CL + 5-item GRS), rated in a fully-crossed g study and d study	Sub-study of objective assessment of TTE technical skills to measure psychomotor competency	S	See above, Neilsen et al.	
			G	- Representative sampling- cases of varying difficulty - Reliability/generalizability analyses – p (physician) x r (raters) x c (cases). For GRS, only 66.6% variance ascribed to p (true differences in physician performance), w 30% interaction effects, particularly physician of different competency-rater and physician of different competency -case level. However for CL, 88.5% variance ascribed to p w much less interaction, only 7%. In D-study, to reach suitable dependability coefficient for high stakes, using CL need 2 cases with 1 random rater, or for GRS need 4 cases with 3 raters.	Favourable, strong
			E		
			I	- impact on learner, program, society? Systematically investigated multiple factors that simultaneously influence test scores insufficiently represented in classical test theory, which helps devise optimal future	Favourable, strong

				test strategy and is sensitive to resource constraints “an informed way out of unreliability” - standard setting? D-study phi coeff >0.9 for high stakes	
Matyal et al. J Cardiothorac Vasc Anesth 2015 [38]	<u>Cardio</u> 11 novices assessed by intraop TEE in sim (first test after intro) then patient (2 nd test after curriculum). Educational intervention: web modules, didactic, supervised sim, pre- and post-test 54 MCQ	Observed kinematic assessment measures TEE image acquisition proficiency. ** educational efficacy study, not assessment validity study	S	- Scoring rubric? 10 cut planes selected by investigator expert consensus based on reproducibility on sim. Kinematic measures: total time, path length, probe accelerations from rest, time-distance integral - Observation format? video - Rater selection/training? Kinematic expert - Rater consistency/accuracy? Blinded analysis of video data novice/expert. Sim motion metrics highly reproducible. - Qualitative? – narratives on training program in questionnaire, not assessment	Favourable, moderate
			G	- Representative sampling? 10 views, kinematic analyses. Different levels of difficulty real world pts *limitn - Reliability/generalizability analyses?	Favourable, moderate
			E	- Real-world correlation? Kinematic metrics (probe accelerations from rest, total time to view, path length, time distance integral) improved from start to end of course. Novices able to perform exam without instructor assistance – clinical transferability. - Scores discriminate groups? Comparing early to after course, fewer image transitions, shorter time to acquire each view.	Favourable, strong
			I	- impact on learner, program, society? May be able to incorporate self-study and time shift activities to comply w duty hour limits. May be able to establish a trainees readiness to perform – manual dexterity may reduce learning curve and enhance the quality of educational experience.	Favourable, moderate
Arntfield et al. Critical Ultrasound Journal 2015 [73]	<u>ER</u> 12 novices assessed by TEE (4 views graded for acceptability) on sim after workshop, and at 6w. Educational intervention: didactic, sim workshop.	Observed TEE assessment measures procedural competence after educational intervention ** educational efficacy study, not	S	- Scoring rubric? Focused scanning protocol for cardiac US in ED by interdisciplinary agreement, 8/28 views then selected 4 that were most relevant. Graded as “acceptability” – subj *limitn - Observation format? video - Rater selection/training? 3 echo experts, no rater training (experts based rating on expert appreciation) *limitn - Rater consistency/accuracy? Blinded, 2 independent reviewers (3rd for consensus if disagreement). IRR fair 0.61-0.8 - Qualitative – narratives via survey re: perceived barriers, comfort but all related to educational intervention not testing	Favourable, moderate

		assessment validity study	G	- Representative sampling? 4 views relevant to ER assessed acceptability of images as “successful image acquisition” not described further, pathological states needed identification - Reliability/generalizability analyses? n/a	Favourable, weak
			E	- Real-world correlation? High levels of success at baseline (82%) and improvement at retention test (96%). At intermediate skill assessment and retention test, 100% pathologic conditions identified. - Scores discriminate groups? Homogeneous group	Favourable, moderate
			I	- impact on learner, program, society? Described some perceived barriers to TEE such as access to probe and how improved confidence arose from curriculum but this is educational intervention, not assessment related - standard setting?	Favourable, weak
Ho et al. Teach Learn Med 2015 [74]	<u>Medical Students</u> 133 students assessed by hand-held ultrasound on pt previously done by expert supervisor twice in 2 wks. Educational intervention: intro, 2w anesthesia rotation, Practical session, bedside scan	Observed HHU testing measures image acquisition skill after an educational intervention ** educational efficacy study, not assessment validity study	S	- Scoring rubric? Success in obtaining 9 basic exams that teach 4 views, maps to curriculum objectives but no reference/rationale - Observation format? - Rater selection/training? Unclear if supervisor rated *limitn - Rater consistency/accuracy? Not given, potential bias *limitn - Qualitative? Narrative – related to the experience of portable TTE not related to assessment, include raters informal feedback which is unique	Favourable, weak but unique aspect
			G	- Representative sampling? Unclear how “determined objectives, how rated appropriately acquired, differing levels of case difficulty. - Reliability/generalizability analyses? n/a	Poor
			E	- Real-world correlation? Inconsistent: difference in student performance for A4C and wk 1 and wk2. implausible - Scores discriminate groups? Expert tutors had 93% success c/w 82% for students, however in MR/MS and PLx in week 2	Inconsistent, weak
			I	- impact on learner, program, society? - standard settings	n/a
Millington et al. J Ultrasound Med 2016 [62]	<u>ER/Critical Care</u> 12 residents with min <30 POCUS experience produced 12 scans, assessed by RACE (5 views assessed for	An objective assessment of point of care echo skills (“RACE” Rapid assessment of competency in echocardiography)	S	- Scoring rubric? Structured, focused interviews w subject experts, video conference, ACCP and Vienna consensus - Observation format? video - Rater selection/training? Video conference after first 10 scored to review each and improve tool, standardize criteria to judge - Rater consistency/accuracy? Independently assessed by 2 experts, deidentified clips, randomized, IRR by Cronbach alpha 0.789. Note: tool	Favourable w some inconsistency, strong

	image generation 6-item scale, and binary image interpretation for 4 aspects, ttl 9 items).	measures point of care echo competency.		image generation good agreement w Cronbach 0.87 but image interpretation poor 0.557 *limitn - Qualitative? Tool designed with emphasis on standard setting, defining competency, thematic saturation to define comprehensive list of features and dimensions, obstacles. Rater feedback modified tool after pilot of rating 10.	
			G	- Representative sampling? Selected image generation and image interpretation 9 items assessed. Strong positive correlations within measures of image generation and within image interpretation but not between two, thus measuring different constructs. However those that didn't correlate could be predicted a prior theoretically. Could not do factor analysis bc limited data set. - Reliability/generalizability analyses? n/a	Favourable, weak
			E	- Real-world correlation? Would have liked to do correlation w OSCE/miniCEX to examine relationship w other variables - Scores discriminate groups? Compared RACE scores early in training (sessions 1-10w) vs late in training (sessions 26-35w) and score lower in early than late stages, with no evidence of itsemf improving at different rates over early to later stages	Favourable, moderate
			I	- impact on learner, program, society? Small study proposes that RACE scale could discriminate between early vs late learners in POCUS – demonstrate in training gains, propose comparing ratings w consequences on educators and learners but do not provide evidence	Favourable, weak
GENERAL/TRAUMA/OBGYN US					
Markowitz et al. J Ultrasound Med 2011 [75]	<u>ER</u> 65 residents (24 PGY1, 16 PGY2, 11 PGY3, 12 PGY4, 2 US fellows) assessed by web-based 41 MCQ re FAST (focused assessment w sono in trauma) exam performance and interpretation	A web-based MCQ assessment tool measures competency in interpretation and clinical correlation of images from EFAST examination in emergency situations	S	- Scoring rubric? NBME guidelines acc to 6 major categories, acceptable to consensus of panel of 5 experts - Observation format? MCQ web-based, not observed - Rater selection/training? MCQ, n/a - Rater consistency/accuracy? MCQ n/a	Favourable, moderate
			G	- Representative sampling? 6 major categories (liver, spleen, pelvic, pericardial, thoracic pneumo, thoracic effusion) w mix of interpretation/assessment, pitfalls; mix of difficulty levels, mix of response format single answer, distractors, still and video. Content experts considered representative and appropriate - Reliability/generalizability analyses? n/a	Favourable, moderate
			E	- Real-world correlation? Scores significantly higher amongst those who had met ACEP guidelines (US course + 25 EFAST exams) criterion validity	Favourable w some

				<p>although still had participants who met ACEP guidelines but didn't get >70% - is this best criterion; of those who had not done a rotation (PGY1+2) those w no rotation significantly worse scores than those doing rotation, with no difference by PGY level bw the no rotation group or rotation group; scores progressively higher the more US exams done and the more EFAST exams done - Scores discriminate groups? Scores progressively higher with level of training by PGY.</p>	inconsistency, strong
			I	<p>- impact on learner, program, society? Helpful as residencies move away from numerical requirements towards competency requirements. Proficiency seen beyond minimum standard important finding</p>	Favourable, weak
Hofer et al. Ultraschall in Med 2011 [76]	<u>Medical students and Residents</u> Most recent 626 final exam OSCE results (300 residents, 326 med students) after 10wk UME or 3d PGME abdominal US course (14 hands-on stations @5m w 1.5m feedback w station CL and GRS 9-item, and 13 diagram stations)	An observed standardized clinical exam measures practical abdominal ultrasound skills after training course ** combined educational efficacy and assessment validity study	S	<p>- Scoring rubric? Catalog of goals and objectives defining which hands-on skills and knowledge to test, designed task sheets, checklists and scoring instructions and adjusted to level possible to master by 90% of 800 med students w 5m/task - Observation format? direct observation - Rater selection/training? Formal training program for instructors (12/18 in present assessment team) incl 1/yr video supported role playing and feedback, all examiners scores compared and discussed until variation <8%, corrective feedback training - Rater consistency/accuracy? Experienced raters more reliable</p>	Favourable, strong
			G	<p>- Representative sampling? 14 hands-on stations and 13 diagram stations. Authors unclear if choice of topics adequately assesses core competencies for abdo US US and Delphi suggested *limitn - Reliability/generalizability analyses? Cronbach alpha 0.69 w 3 hands on and 2 diagram stations; >0.8 w 8+stations;</p>	Favourable, moderate
			E	<p>- Real-world correlation? - Scores discriminate groups? discrimination coefficients of 14 hands-on stations very high (0.48) discriminating overall outstanding examinees and poor performers but lower than expected for diagram stn 0.16. Very large sample studied.</p>	Favourable, strong
			I	<p>- impact on learner, program, society? First comprehensive OSCE w high values of discrimination coefficients useful for high stakes exam. Cut off values can be used to define pass fail</p>	Favourable, strong
Thoirs et al. Aust J Educ Tech 2012 [77]	<u>Sonographer Students</u> 5 novices assessed by pre and post	Observed MSK US testing measures psychomotor procedural	S	<p>- Scoring rubric? Baseline competency determined levels of difficulty across groups – no a priori blueprint *limitn. - Observation format? direct observation - Rater selection/training? Accredited experienced sonographer</p>	Favourable, weak

	competency testing on a pt (Demonstrate 17 anatomic structures). Educational intervention: DVD instructional tool and supervised and independent clinical practice sessions.	competence after a teaching intervention. ** educational efficacy study, not assessment validity study		- Rater consistency/accuracy? Single rater - Qualitative – At 3mo, structured interviews re: instructional delivery but not assessment testing itself * limitn	
			G	- Representative sampling? Comprehensive nont sampled (17 anatomic structures in ankle on DVD); difficulty determined by participant baseline competency; did not assess other aspects - Reliability/generalizability analyses? n/a	Favourable, weak
			E	- Real-world correlation? no control. Post training, competency improved for moderately-difficult structures and somewhat for difficult structures but still quite difficult, though overall scores improved. Not clear on statistical significance. Greatest improvement in those w low and moderate baseline competence - Scores discriminate groups? n/a, homogeneous group.	Favourable, weak
			I	impact on learner, program, society? - standard setting?	n/a
Madsen et al. Ultrasound Ob Gyn 2014 [78]	<u>OB/GYN</u> 16 novices and 12 faculty assessed on sim x two iterations of same 7 modules with metrics by sim. After 2 mos, educational intervention for novices to reach score for experts w feedback. Pilot on 3 med students, 3 residents and 1 consultant using 14 modules.	Observed assessment with sim metrics measures transvaginal ultrasound performance ** combined educational efficacy and assessment validity study	S	- Scoring rubric? Pilot participant comments informed scoring rubric/selection and 153 manufacturer's defined metrics defined w no description of how selected "- automated, dichotomous *limitn - Observation format? Sim automated - Rater selection/training? n/a – sim, no subjective bias, reproducible (test/re-test high, ICC 0.93). - Rater consistency/accuracy? n/a - sim	Favourable, weak
			G	- Representative sampling? Selected 7 of 14 possible modules – unclear if representative. Only 48 selected metrics discriminated expertise - Reliability/generalizability analyses? n/a	Favourable, weak
			E	- Real-world correlation? Experts took less time to complete initial 2 testing iterations - Scores discriminate groups? Experts scored significantly higher than novices on 48 metrics w $p < 0.05$ and overall score	Favourable, strong
			I	impact on learner, program, society? First study to explore US learning curves using valid, reliable metrics - standard setting – criterion based training * able to define a pass/fail level via contrasting groups and novices worked to get to that point where they have automaticity and "fit for supervised clinical practice"	Favourable, strong
Jaffer et al. Heart Lung	<u>Vascular Surgery</u>	A modified objective	S	- Scoring rubric? Derivation of OSATS per domains of content analysis using conditions of learning w 2 experts informally agreeing, scoring arbitrarily	Favourable, strong

Vessels 2014 [79]	9 novices, 8 intermediate and 6 experienced (med students, residents, sonographer students) participants assessed on modified DUOSATS (5-item GRS and 9 domain CL) on simulated 70% stenosis model.	assessment tool (mDUOSATS) measures duplex arterial stenosis detection procedural skills.		assigned to reflect progression from lower to higher order concepts *limitn. GRS (5 pt) and checklist - Observation format? video review - Rater selection/training? 4 experts - Rater consistency/accuracy? Blinded to identity IRR high cronbach alpha CL 0.97, GRS 0.96	
			G	- Representative sampling? 11 domains selected based on essential features by content analysis - Reliability/generalizability analyses?	
			E	- Real-world correlation? correlation w GRS and DUOSATS score thus same trait measured? - Scores discriminate groups? Novice – Intermed – Experts for GRS trend not significant; for avg score of CL+GRS no significant difference either until change expertise level definitions to reflect prev stenosis measurement experience. Of 9 items in CL, 4 significant. *limtn, sens to spec skills related to skill of interest	Inconsistent, moderate
			I	- impact on learner, program, society? Suggest follow individual domains of significance when using DUOSATS in formative assessment, with carefull feedback here to focus and hasten training - standard setting? *not subject to halo effect, ROC modelling cut-point using scores of those experienced in detection – AUC 0.895, determined high specificity trading off lower sensitivity (less experienced unlikely to be considered competent though more experienced may not achieve competence)	Favourable, strong
Tolsgaard et al. US OB GYN 2014 [80]	OB/GYN 10 novice, 10 residents and 10 experts in OB/GYN US assessed scanning transabdominal fetal biometry or transvaginal systematic pelvic scan on 30 different pts by OSAUS (5 domains, 5-item GRS)	An objective assessment of Transvaginal/transabdominal US skills (mOSAUS) measures procedural competence.	S	- Scoring rubric? Modified OSAUS SCALE previously described in literature, derived by multispecialty Delphi(Tolsgaard 2013) emphasis on equal weighting of components - Observation format? video of hand movements and US output - Rater selection/training? 2 consultants, blinded, anonymized scans and audio distortion. 4 videos rated prior individually and then discussed to consensus, - Rater consistency/accuracy? ICC 0.89 inter-rater;	Favourable, strong
			G	- Representative sampling? 5 Equal weighted 5 items (applied knowledge of US equipment, image optimization, systematic exam, interpret of images, document'n of exam) selected of possible 7 since more appropriate for specific instructions in this study. Rationale discussed in (Tolsgaard 2013). Note: different levels of difficulty of pts not studied as interaction/source of potential error *limitn. tool consistent Cronbach alpha 0.96	Favourable, moderate

			E	<ul style="list-style-type: none"> - Real-world correlation? time for fetal biometry decreased with expertise level, n/s difference w systematic exam * - Scores discriminate groups? Mean score significant difference bw 3 groups in fetal biometry exams and pelvic exams. Significant difference bw novice-intermed and intermed-senior. However looking at data individual components of score (particularly domain 2 and 5) very wide SD and not clear if differentiates 	Favourable with some inconsistency, strong
			I	<ul style="list-style-type: none"> - impact on learner, program, society? established credible pass/fail a benchmarks score set to allow for future meaningful use. Suggested considering CUSUM scores currently used to detect suboptimal performance by sonographers on scans though not done. Note: OSAUS had content validity prev but this added expertise discrimination thus useful - standard setting? Pass/fail using contrasting groups method (non-competent c/w competent performers to determine best discrimination bw groups) – at cut point, all novices failed (no false pos) and all seniors passed (no false neg). Interestingly, cutpoint score differs for TV vs TAB US and intermed passed more TV, perhaps different skill proficiency from experience? 	Favourable, strong
Jaffer et al V-DUOSATS 2015 [81]	<u>Vascular Surgery</u> 24 participants divided into 4 groups (8 novice, 2 junior, 2 intermed, 2 senior, 10 expert) assessed on modified V-DUOSATS (4-item GRS and 6 domain CL) assessed on simulated model of venous reflux.	A modified objective assessment tool (V-DUOSATS) measures duplex US venous reflux detection procedural skills.	S	<ul style="list-style-type: none"> - Scoring rubric? Derivation of OSATS per domains of content analysis using conditions of learning w 2 experts informally agreeing, scoring arbitrarily assigned to reflect progression from lower to higher order concepts *limitn. - Observation format? video - Rater selection/training? 3 blinded expert; also 5 novice assessors w min US training and no experience independently assessed - Rater consistency/accuracy? IRR high with cronbach alpha 0.8, 0.82. Novice raters cw expert raters correlation, R0.5 w no bias 	Favourable strong
			G	<ul style="list-style-type: none"> - Representative sampling? 6 of 8 domains thought relevant for simulation. Note: correlation bw D-DUOSATS score and GRS – same trait measured? *limitn. - Reliability/generalizability analyses? n/a 	Favourable, moderate
			E	<ul style="list-style-type: none"> - Real-world correlation? V-DUOSATS score negatively correlated with % error in reflux time estimation – “end product style assessment” however GRS did not - Scores discriminate groups? Sign differences in both V-DUOSATS and GRS scores across 4 groups of experience using US experience, duplex US experience and reflux time experience 	Favourable, strong

			I	<ul style="list-style-type: none"> - impact on learner, program, society? Propose use in formative assessment, can focus on those domains that seem to differentiate bw the different experience groups (see Table 2) - standard setting? ROC plotted for expert group AUC 0.88, sens and spec cutpoint made to maximize spec (no incompetent operators considered compent) 	Favourable, strong
Todsén et al. Annals of Surgery 2015 [82]	ER/Gen Surg 12 novice, 8 intermed, 2 expert POCUS users assessed by OSAUS (exam of 4 simulated cases @ 5 domains per case GRS 5-item scale)	An objective assessment of abdominal POCUS skills (mOSAUS) measures procedural technical competence.	S	<ul style="list-style-type: none"> - Scoring rubric? Modified OSAUS SCALE previously described in literature, derived by multispecialty Delphi (Tolsgaard 2013) emphasis on equal weighting of components - Observation format? video, US screen output merged - Rater selection/training? 2 radiologists subspec in US/interventions. 90m training session after data acquired w 5 pilot videos reviewed and discussed until consensus - Rater consistency/accuracy? Blinded, independently assessed, anonymized clips, data electronically transferred minimizing error. IRR in g-study very low, 5.75% of variance in scores 	Favourable, strong
			G	<ul style="list-style-type: none"> - Representative sampling? Equal weighted 5 items (applied knowledge of US equipment, image optimization, systematic exam, interpret of images, document'n of exam) selected of possible 7 since more appropriate this study. Rationale in (Tolsgaard 2013). Cases selected where gen surg would POCUS in ER. Same 4 pts scanned by all. - Reliability/generalizability analyses? Highest source of variance in scores was physicians (44%) with interactions bw case and assessor substantial (24%) and physician/case and assessor (22%). 	Favourable, strong
			E	<ul style="list-style-type: none"> - Real-world correlation? Strong correlation bw OSAUS score and number of sonographically verifiable correct diagnoses. - Scores discriminate groups? Yes using group comparisons: mean scores significantly higher in experts c/w intermediate and novices. G-theory study true variance = differences in OSAUS scores bc of different competence bw individuals. 44% of variance from physicians but substantial interaction effects from assessor and case (which applied across the board to all participants), and physician, case and assessor. D-study predicted 5 ratings from 1 assessor ensures gen co-eff >0.8 	Inconsistent, strong
			I	<ul style="list-style-type: none"> - impact on learner, program, society? Generalizability co-eff >0.8 for high stakes exams and >0.6 for formative exams; here was 0.81 (81% of score d/t true score not error of measurements). 	Favourable, moderate

Ziesmann et al. J Trauma Acute Care Surg 2015 [83]	<u>Trauma/Gen Surg</u> 12 novice, 12 experts assessed doing Focused Assessment with Sono for Trauma US on 1 pt by QUICK model (8-domain 5-item GRS, 24-item CL binary y/n)	An objective measurement of psychomotor skills measures FAST (Trauma) US competence.	S	- Scoring rubric? Derivation of OSATS with scale and CL content developed through Delphi with 10 experts. - Observation format? video - Rater selection/training? 2 FAST experts, oriented, 4 out of sample videos scored for consensus forming - Rater consistency/accuracy? Blinded, independently scored, IRR Kappa for CL 0.79, GRS 0.61 (moderate-substantial)	Favourable, strong
			G	- Representative sampling? GRS domains defined by Delphi, removed 1 domain as all autonomous in study. CL component 24 anatomic landmarks in the 4 regions required in FAST scans. patient scanned x 24 times, single level of difficulty *limtn. - Reliability/generalizability analyses? mean squared error s bw observed and predicted scores for TSC 0.28, GRS 0.08	Favourable, moderate
			E	- Real-world correlation? n/a - Scores discriminate groups? Experts significantly higher scores than novices for ttl CL score; none: for ¾ anatomic regions but for pelvic it was not seen thus inconsistent and may not detect diff vs underpowered study *limitn; experts significantly better GRS and all individual domains.	Favorable with some inconsistency, Moderate
			I	- impact on learner, program, society? Hypothesis generating but “a first step” - standard setting? Delphi set standard expected for sonographer to meet expectations of safe practice. Univariate predictor of expert status modelling using ROC for CL sens 86% and spec 75%, AUC 90%. ROC for GRS Sens 93%, spec 92%, AUC 98%.	favorable, weak
Ziesmann et al. J Trauma Acute Care Surg 2015 [84]	<u>Trauma/Gen Surg</u> 12 novice, 12 experts assessed doing Focused Assessment with Sono for Trauma US on 1 pt by hand motion analysis with affixed magnet as scanning	Objective measurement of sim-based hand motion metrics assesses FAST ultrasound competence	S	- Scoring rubric? HMA metrics. Not clear how selected. - Observation format? fully automated - Rater selection/training? n/a automated - Rater consistency/accuracy? n/a automated	Favorable, weak
			G	- Representative sampling? Anatomic locations of FAST scan are pericardia, peritoneal. Time, # movements, path length travelled are automated measurements reflecting efficiency. However did not discuss possibilities not included or rationale - Reliability/generalizability analyses? n/z	Favorable, weak
			E	- Real-world correlation? criterion referenced to QUICK score, matched with HMA outcome. Negative correlation bw QUICK and HMA bc as QUICK scores improve. Values -0.18 to -0.6 suggest measuring different traits complementary, reflecting knowledge and technical efficiency	Favorable, strong

				- Scores discriminate groups? Novices take longer for total path length travelled, and many more movements (less automatic). Surprisingly, time was n/s b/w groups but suspected d/t pt factors	
			I	- impact on learner, program, society? First to use HMA for FAST US. Opportunity for real-time feedback, minimizing local HR reliance for training (automated), serially assess using CUSUM to map learning curve - standard setting – modelled ROC curves assuming a path length differentiating expert from novice performance	Favorable, moderate
Chaudery et al. J Surg Education 2015 [85]	<u>Trauma/Gen Surg</u> 10 novices, 10 intermediates, 11 experts in Focused Assessment with Sono for Trauma US assessed on FAST sim measuring time based measures, post-study. questionnaire	Objective assessment of time-based metrics in FAST scan measures FAST US performance	S	- Scoring rubric? – tool for FAST assessment exists, thought by authors to be too subjective thus developed new tool with time based measures but how selected not discussed - Observation format? sim - Rater selection/training? 2 radiology experts - Rater consistency/accuracy? Blinded to assignment group, 2 independent, blinded radiologists w kappa 0.72 - Qualitative – questionnaire re: usefulness, realism of simulator but nothing re: assessment testing	Favorable, weak
			G	- Representative sampling? Not clear how selected time-based measures reflective of sampling of test universe. - Reliability/generalizability analyses? n/a	Favorable, weak
			E	- Real-world correlation? - Scores discriminate groups? Difference between novice-intermed-experts time to scan; time to identify abN; n/s difference in time to freeze best image * also n/s scan total between novice-intermediate though intermed-experts and novice-experts seen. *limitn	Favorable, some inconsistency, moderate
			I	- impact on learner, program, society? “potential to accelerate novices up the learning curve” hypothesis generation - standard setting?	Favorable, weak
Patrawalla et al. J Grad Med Ed 2015 [86]	<u>Critical Care</u> 28 fellows 1 year post US training course assessed by CCUS assessment on DVT: 12 step dichotomous CL with 3-item GRS and echo: 12 step CL and 3-item GRS for 2	An objective assessment of CCUS measures procedural competence for DVT and echo LV function.	S	- Scoring rubric? Modified Delphi expert panel of 4, procedure specific CL items for 2 US scenarios using literature and guidelines - Observation format? live and video - Rater selection/training? faculty w extensive experience in CCUS application and education, Delphi panel members; scripted instructions and feedback using structured report card *limitn no rater guide and no behavioral anchors for incorrect task performance - Rater consistency/accuracy? 1 rater rated all and 2 nd rater did subset of 10. Cronbach alpha for DVT CL 0.85, for echo 0.92. Kappas for DVT .21-1 live vs video; 0-0.62 between video; 100% agreement GRS. For Echo	Favorable, moderate

	echo views. MCQ also done.			kappas 0.29-0.58 live vs video; 0.74-1 bw 2 video; GRS 0.44 Lax, 0.58 * overall – poor IRR w no correlation, video footage inadequate limitn	
			G	- Representative sampling? Imaged same healthy actor. Sampling of tasks selected systematically by Delphi to consensus. - Reliability/generalizability analyses? n/a	Favorable, moderate
			E	- Real-world correlation? n/a - Scores discriminate groups? Homogeneous group, no data	n/a
			I	- impact on learner, program, society? Propose remote asynchronous viewing as an alternative to time intensive direct supervision	Favorable, weak
Schmidt et al West J Emerg 2016 [87]	<u>ER</u> 9 experienced senior residents assessed on an POCUS OSCE for image acquisition using standardized patient (5 stations each had CL with bw 7-18 items dichotomous) and image interpretation by computer MCQ video quiz	A POCUS OSCE objectively measures US acquisition competence.	S	- Scoring rubric? OSCE style, 5 stations represented 5 core US skills FAST/.aorta/echo (15pts)/pelvic/central line. CLs for 4/5 domains created by Academy of ER US, CVC prev published. - Observation format? direct observation - Rater selection/training? 2 raters - Rater consistency/accuracy? Non-blinded raters for image acquisition *limitn - bias	Favorable, weak
			G	- Representative sampling? Pre-defined domains on CL per Academy of ER CORD defined skills or CVC literature. Included equally weighted core and advanced not explaining how many items indicate competency. Of the 5 stations, variability in avg scores, and when excluding advanced US competencies, the avg total score increased. Single standardized patient. Small cohort. - Reliability/generalizability analyses? n/a	Favorable, moderate
			E	- Real-world correlation? No indication on OSCE per CORD Academy of ER, how scores correlate with clinical performance. - Scores discriminate groups?	n/a
			I	-- impact on learner, program, society? - standard setting? Unclear as to what score represents competency.	Variable, weak
Dyre et al. Utraschall in Med 2016 [88]	<u>OB/GYN</u> 20 novices and 9 experts assessed on 10 modules by 126 simulator metrics (dichotomous). The significant	An objective assessment of abdominal US using simulation-based metrics measures US competence	S	- Scoring rubric? 10 modules selected by 2 experts, metrics based on module's relevance to basic OB US (established validity evidence in ISUOG guidelines 2014). For final sim test, only discriminating metrics (40 of 126). - Observation format? automated sim - Rater selection/training? n/a sim - Rater consistency/accuracy? n/a sim – reproducible. Test/retest ICC 0.62 for 5 novices reaching mastery twice.	Favorable, moderate

	discriminating metrics were used on subsequent testing of novices with instructor feedback after each module re failed metrics.	and as a target for mastery learning ** combined educational efficacy study and assessment validity study	G	- Representative sampling? 10 cases selected by 2 experts as relevant; domains per ISUOG. Since automated, image interpretation and med mgmt. not included *limitn - Reliability/generalizability analyses? n/a	Favorable, strong
			E	- Real-world correlation? shorter median time to complete test for expert vs novice, larger variation in clinically impmt measurements cw experts though n/s for 2 of 3 measurements. Unclear if transfers to clinical *limitn - Scores discriminate groups? Overall scores discriminated novice and metric, but only 32% (40/126) of metrics discriminated novice vs expert. *	Favorable with some inconsistency, moderate
			I	- impact on learner, program, society? mastery learning level established 90%, median score of metrics w validity evidence in expert group, learning curve plateau seen if achieved twice - standard setting? Pass/fail established by contrasting groups method dividing experts from novces at 72%	Favorable, strong
Amini et al. Adv Med Educ Prac 2016 [89]	ER 52 residents PGY 1-3 assessed on single case, sim-based POCUS OSCE. 1d assessment workshop incl OSCE, MCQ, sim-based diagnostics, management, hands-on education station.	A sim-based OSCE assesses POCUS technical skill competency. ** educational efficacy study not assessment validity study. Analysis of OSCE only.	S	- Scoring rubric? Not provided, not pilot tested, not validated prior to implementation - Observation format? not described - Rater selection/training? not described - Rater consistency/accuracy? Not described	Insufficient, weak
			G	- Representative sampling? Single case of hypotension selected, unclear re: rubric and unclear re: representative sampling of domains. - Reliability/generalizability analyses?	Insufficient, weak
			E	- Real-world correlation? - Scores discriminate groups? OSCE scores higher PGY 3 vs PGY 1 but PGY 2 lowest ?limit'n plausibility vs proximity to US rotation as interns	Inconsistent, weak
			I	- impact on learner, program, society?	insufficient
Black H et al. Cureus 2016 [90]	<u>Medical Students</u> POCUS OSATS assessment for medical students (CLs with binary response for 9 items Aorta, 11 item subxiphoid cardiac, 11 item focused abdo and 9 domain GRS w 5-item scale).	A POCUS assessment tool for UME curriculum assesses proficiency over time.	S	- Scoring rubric? OSATS derived, Modified Delphi with non-purposive sampling of expert panel of 18 ER POCUS experts, systematic process. Milestones for US suggested in guidelines for medical students (2016). - Observation format? n/a - Rater selection/training? n/a - Rater consistency/accuracy? n/a	Favorable, moderate
			G	- Representative sampling? Purposive sampling - Reliability/generalizability analyses?	n/a
			E	- Real-world correlation? - Scores discriminate groups?	n/a
			I	- impact on learner, program, society?	n/a

APPENDIX G – Data Collection Form



V3 Revised July 4, 2016



Data Collection Form

STUDY ID # _____

Level of Training (as of July 2016): ☐ C1 ☐ C2 ☐ C3 ☐ >=C4Sex: ☐ M ☐ F

Previous Echo seen: _____

Previous complete Echo performed: _____

Previous complete Echo interpreted: _____

Previous non-cardiac ultrasound seen: _____

Previous complete non-cardiac ultrasound performed: _____

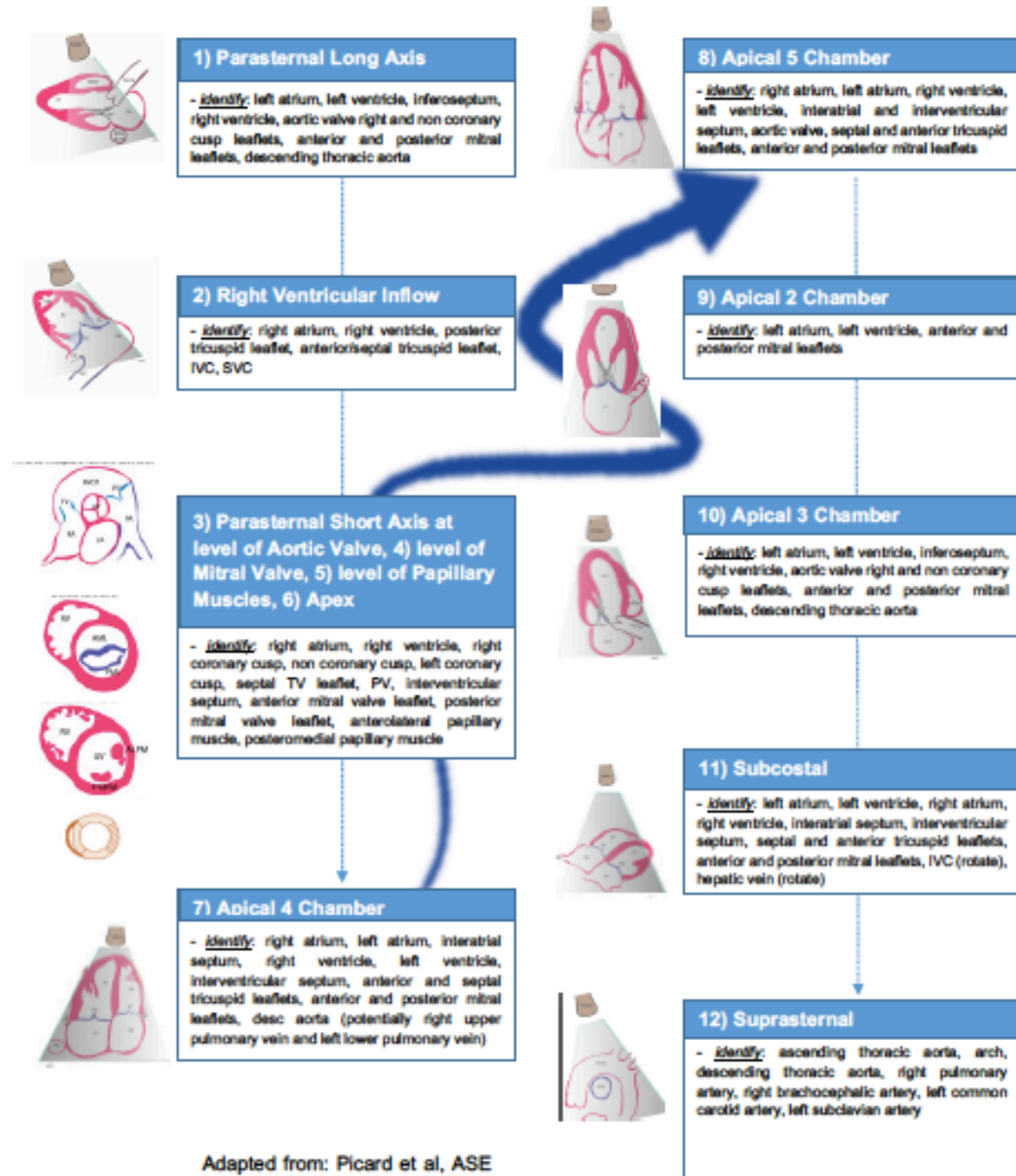
Previous complete non-cardiac ultrasound interpreted: _____

Cell/Phone #: _____

A member of the research team not involved with rating will contact you for a brief minute phone interview 2 weeks after today, and this number will then be destroyed and not included in our data collection

APPENDIX H – Standardized Echocardiography Scanning Template

TTE Standard Scanning Protocol



APPENDIX I – Echocardiography Competence Assessment Tool

Trainee #:	Date:	Rater:	Direct <input type="checkbox"/> Off-line <input type="checkbox"/>
------------	-------	--------	---

The purpose of this scale is to assess the trainee's ability to scan a patient.
Please use the PGY Competence-aligned scale to evaluate each echocardiography view

EXPECTED PERFORMANCE OF PGY 4 = competence expected early in cardiology training

EXPECTED PERFORMANCE OF PGY 5-6 = competence expected advancing in cardiology training

EXPECTED PERFORMANCE AT END OF PGY 6 = competence expected at the end of training

	Not Done (0)	Expected performance of PGY 4 (1 point)	Expected performance of PGY 5-6 (2 points)	Expected performance at End of PGY 6 (3 points)
PARASTERNAL L AXIS				
<u>Use of screen</u> (width and depth) - image bottom: IL wall, top: RVOT, left: mid LV, right: LVOT; IL wall parallel to beam	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<u>Key structures:</u> (10) LA, LV, AMVL, PMVL, RCC, NCC, DA, LVOT, RV, IVS	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
RV INFLOW				
<u>Use of screen</u> (width and depth) Image bounded at bottom: RA, top: RV, left: RVFW, right: RV	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<u>Key structures:</u> (6) RA, RV, PTVL, ATVL/STVL, IVC, SVC	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
PARASTERNAL S AXIS @ AORTIC VALVE				
<u>Use of screen:</u> (width and depth) Image bounded at bottom: RA/LA, top: RVOT, left: RV, right: PA	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<u>Key structures:</u> (10) RA, LA, STVL, RV, IAS, RCC, NCC, LCC, PV, MPA	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
PARASTERNAL S AXIS @ MITRAL VALVE				
<u>Use of screen:</u> (width and depth) Image bounded at bottom: LV, top: RV, left: RV, right: LV	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<u>Key structures:</u> (5) RV, AMVL, PMVL, LV, IVS	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
PARASTERNAL S AXIS @ PAP MUSCLES				

<i>Use of screen:</i> (width and depth) Image bounded at bottom: LV, top: RV, left: RV, right: LV	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (5) RV, LV, IVS, ALPM, PMPM	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
PARASTERNAL S AXIS @ APEX				
<i>Use of screen:</i> (width and depth) Image occupied by LV w a thin crescent of RV	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (1) LV	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
APICAL 4 CHAMBER				
<i>Use of screen:</i> (width and depth) Image bounded at bottom: RA and LA, top: LV apex, left: RV and RA, right: LV and LA	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (11) RA, LA, IAS, RV, LV, IVS, ATVL, STVL, AMVL, PMVL, DA, (RUPV, LLPV)	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
APICAL 5 CHAMBER				
<i>Use of screen:</i> (width and depth) Image bounded at bottom: RA and LA, top: LV apex, left: RV and RA, right: LA and LV	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (11) RA, LA, IAS, RV, LV, IVS, AoV, STVL, ATVL, AMVL, PMVL (RUPV)	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
APICAL 2 CHAMBER				
<i>Use of screen:</i> (width and depth) Image bounded at bottom: LA, top: LV apex, left: LV right: IVS	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (4) LA, LV, AMVL, PMVL	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
APICAL 3 CHAMBER				
<i>Use of screen:</i> (width and depth) Image bounded at bottom: LA, top: LV apex, left: LV right: IVS and LVOT	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (10) LA, LV, AMVL, PMVL, RCC, NCC, DA, LVOT, RV, IVS	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
SUBCOSTAL WINDOW				
<i>Use of screen:</i> (width and depth) Image bounded at bottom: LV, top: RV, left: RV, right: LV	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)

<i>Key structures:</i> (11) IVC, RA, RV, LA, LV, IVS, IAS, ATVL, STVL, AMVL, PMVL (hep vein)	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
SUPRASTERNAL WINDOW				
<i>Use of screen:</i> (width and depth) Image top: aortic arch, left: prox asc aorta, desc thoracic aorta	Did not adjust screen parameters (width and depth)	Almost able to adjust a screen parameter (width OR depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth)
<i>Key structures:</i> (7) Proximal ascending aorta, Arch, descending thoracic aorta, RPA, R brachiocephalic, L common carotid, L subclavian	Not able to identify any key structures	Identification of SOME (<75%) key structures	Identification of MOST (75-99%) key structures	Identification of ALL (100%) key structures
Total Score (/72)				

Global Impression


Is this study of diagnostic quality? YES ☐ NO ☐

Feedback: Please provide as much information as possible for participants on specific aspects of the echo they did well. Please provide specific suggestions for improvement and any additional comments.

Signature: Trainee _____

Rater

APPENDIX J – Web-based Echocardiography Competence Assessment Tool for Off-line Rating



EchoSim Competence Assessment Tool

Date assessed

DD

MM

YYYY

Date

/

/

Rater Initials

Trainee ID #

Rating location

☐ off-line
 ☐ real-time

The **purpose** of this tool is to assess **ability to scan**.

Please **evaluate each echo view** for

- ability to optimize width/depth** to see relevant structures
- ability to identify key structures** from the acquired images

PGY-ANCHORS

NOVICE = competence of a novice starting training
 PERFORMANCE AT END OF C1/PGY 4 = competence early in training
 PERFORMANCE AT END OF C2/PGY 5 = competence advancing in training
 PERFORMANCE AT END OF C3/PGY 6 = competence at the end of training

1) PARASTERNAL LONG AXIS

	Novice	Performance end of PGY 4	Performance end of PGY 5	Performance end of PGY 6
use of screen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
structure identification	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

PARASTERNAL LONG AXIS	Novice	End PGY 4	End PGY 5	End PGY 6
use of screen image bounds: bottom: IL wall, top: RVOT, left: mid LV, right: LVOT; IL wall parallel to beam	Did NOT adjust screen parameters (width or depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth) but needs optimization	Able to adjust both screen parameters (width AND depth) without requiring further optimization
identify structures (10) LA, LV, AMVL, PMVL, RCC, NCC, DA, LVOT, RV, IVS (only 1 point for "MV"; "AoV", if no DA in image mark out of 9; bonus coronary sinus)	Able to identify FEW (<50%) key structures	Identification of SOME (50-75%) key structures	Identification of MOST (76-99%) key structures	Identification of ALL (100%) key structures

2) RIGHT VENTRICULAR INFLOW

	Novice	Performance end of PGY 4	Performance end of PGY 5	Performance end of PGY 6
use of screen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
structure identification	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

RIGHT VENTRICULAR INFLOW	Novice	End PGY 4	End PGY 5	End PGY 6
use of screen image bounds: bottom: IL wall, top: RVOT, left: mid LV, right: LVOT; IL wall parallel to beam	Did NOT adjust screen parameters (width or depth)	Able to adjust one screen parameter (width OR depth)	Able to adjust both screen parameters (width AND depth) but needs optimization	Able to adjust both screen parameters (width AND depth) without requiring further optimization
identify structures (10) LA, LV, AMVL, PMVL, RCC, NCC, DA, LVOT, RV, IVS (only 1 point for "MV"; "AoV", if no DA in image mark out of 9; bonus coronary sinus)	Able to identify FEW (<50%) key structures	Identification of SOME (50-75%) key structures	Identification of MOST (76-99%) key structures	Identification of ALL (100%) key structures

[SURVEY PREVIEW MODE] Edit X Sherryn

https://www.surveymonkey.net/r/?sm=ES1pxyWxwIW02147IRiHdzkcs7x3h_2F0wLtTPPZ60FHE_3D

Apps Bookmarks Bar

Global Impression: Was this study of diagnostic quality?


☐ Yes ☐ No

FEEDBACK: Please provide as much information as possible for participants on specific aspects of the echo they did well. Please provide specific suggestions for improvement and any additional comments.

100%

Done

Powered by

 SurveyMonkey®

APPENDIX K – Interview Guides

Participant Interview Guide

Description: Participants who underwent assessment testing were contacted two weeks after session by telephone. The semi-structured interview started with the following questions:

- As you read through the feedback, what conclusions did you draw about your performance?
 - How do those conclusions match with your own impressions you formed on the day of your performance?
- Do you think the feedback will make a difference to how you will perform/read an echocardiography in practice?
 - Prompt: If yes, how will you use it?
 - Prompt: If not, why not? What could be changed about this process to give you that motivation?
- Different people provided your feedback – sonographers and physicians– does knowing this affect the way you interpret the feedback you received?
- How could we make the feedback process better for your learning?
- Take me through the ECAT testing – how was that experience for you? How were you feeling at the end of it?

Rater Interview Guide

Description: Raters were interviewed at the completion of the study. The semi-structured interview started with the following questions:

- Please can you talk me through how you used the ECAT tool to provide formative feedback for the participants?
 - Prompt: how did you go about translate the ratings from the tool into the narrative feedback you provided?
- You provided feedback (in person or offline). What did you find challenging about using the form to provide feedback in that situation?
- How could we make the feedback process better for you as a rater?
- What changes would you make to the tool or any part of the process to facilitate giving the best formative feedback possible?

VITA

NAME: Sherryn Rambihar

EDUCATION: B. Sc. Human Biology & Sociology, University of Toronto, Toronto, Ontario, Canada, 2001.
M.D., Western University, London, Ontario, Canada, 2005.

POSTGRADUATE: Postgraduate Training, Internal Medicine, Western University, London, ON, Canada 2005-2008.

Postgraduate Training, Cardiology, McMaster University, Hamilton, ON, Canada, 2008-2011.

Echocardiography Clinical and Research Fellowship, Mount Sinai Hospital, University of Toronto, Toronto, ON, Canada, 2011.

APPOINTMENTS: Consultant Cardiologist, Cardiac Rehabilitation, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada. (2017 - present)

Cardiology Private Practice, Toronto, ON, Canada (2016 – present)

Courtesy Staff, Division of Cardiology, Mount Sinai Hospital, Toronto, ON, Canada (2012 – present)

Clinical Associate, Division of Cardiology, Department of Medicine, North York General Hospital, Toronto, ON, Canada (2009 – present)

Staff Cardiologist, Division of Cardiology, Department of Medicine, Women's College Hospital, Toronto, ON, Canada (2012-2016).

TEACHING: **Adjunct Assistant Professor, Division of Cardiology, Department of Medicine, University of Toronto**

Undergraduate MD: Small Group Tutor, Year 1 New Foundations Curriculum, Year 2 Mechanisms, Manifestations and Management of Disease Course, Core lecturer to CC3 and CC4 Peters-Boyd Academy Students, Transition to Residency Students, and elective/observer students, OSCE assessor, Professionalism lecturer.

Postgraduate MD: Pericardiocentesis and Introduction to Echocardiography Academic Core Session co-leader, physical examination practice session leader at complex ambulatory care clinic, family medicine

resident lecturer, core lecturer at cardiology resident academic half day, clinical teaching of transesophageal echocardiography to cardiology residents, small group leader at internal medicine half day ECG training.

Interprofessional Education: chair of cardiology interprofessional rounds, echo quality improvement program leader, lecturer at anesthesia assistant academic education day, small group leader for practice-based continuing professional development of family physicians.

Curricular Design: Echo E-Atlas (2011, Division of Cardiology, University of Toronto, Toronto, ON, Canada); Web Medical Education: Complexity and Women's Health (2004, Harvard Medical School, Boston, MA). Curriculum Design: Global Health (2001, Western University, London, ON, Canada).

HONORS:

American College of Cardiology Award for Top Canadian Abstract, Washington, DC, 2017.

Young Clinician Education Award, Women's College Hospital, Toronto, Canada, 2014.

Canadian Society for Clinical Investigation/PSI National Award for Resident Clinical Research, McMaster University, Hamilton, ON, Canada, 2011.

American Heart Association Cardiology Trainee Award for Excellence, Chicago, IL, 2010.

Gold Medal Winner, McMaster Department of Medicine Resident Research Day, McMaster University, Hamilton, ON, Canada, 2010.

Ontario Medical Association Medical Student Achievement Award, Western University, London, ON, Canada, 2003.

Marlene Kirsh Korman Scholarship, University College, University of Toronto, Toronto, ON, Canada, 2001

Reuben Wells Leonard Scholarship, University College, University of Toronto, Toronto, ON, Canada 2000

Dr. James A. and Connie Dickson Scholarship for Sciences and Math, University College, University of Toronto, Toronto, ON, Canada, 1999

Mossman Admissions Scholarship, University College, University of Toronto, Toronto, ON, Canada 1998.

ABSTRACTS:

- 2017 **Rambihar S**, Lineberry M, Nesbitt G et al. Echo Milestones: A Novel simulation-based echocardiography competence tool for formative assessment in cardiology training. *J Am Coll Cardiol* 2017; 69(11): S2514.
- 2014 **Rambihar S**, Khanduja K, Ahmed S, et al. Use of the Modified Delphi technique to develop a competency-based simulation enhanced TTE curriculum for novices. *J Am Soc Echo* 2014; 27(6) B97.
- 2009 White J, Armstrong S, **Rambihar S** et al. Abnormal myocardial perfusion in hypertrophic cardiomyopathy: Preliminary findings of an MRI study. *Journal of Cardiovascular Magnetic Resonance* 2009, 11 (Suppl 1): P55. (Abstract)
- 2009 Abramson B, **Rambihar S**, Jaakimainen L et al. Gender and Regional Variation in Quality of Care of Ischemic Heart Disease in a Large Canadian Cohort. *J Am Coll Cardiol* 2009; (53)10, S1, A367.
- 2004 **Rambihar S**, Raha A, Van Dorp N, Howard J. (R)evolution in medical education: Can MEDS2000 on a global axis. *Clin Invest Med* 2004; 27(4):197, abstract # 81.
- 2000 **Rambihar S**, Woodhouse R, Tenenbaum J et al. Women's Heart Health in the Postgraduate Medical Curriculum, an educational needs analysis. *Ann Int Med* 2000;35(6).
- 2000 **Rambihar SP**, Rambihar VS, Jagdeo DG, et al. Women, Ethnic Variation, and Coronary Artery Disease. *Can J Cardiol* 2000;16(suppl B):49B.

PUBLICATIONS:

- 2015 Bhatia RS, Ivers N... **Rambihar S...** et al. Design and methods of the Echo WISELY (Will Inappropriate Scenarios for Echocardiography Lessen Significantly) study. *Am Heart J* 2015;170(2):202-9.
- 2014 **Rambihar S**, Sanfilippo AS, Zasson Z. Mitral Chordal-Leaflet-Myocardial Interactions in Mitral Valve Prolapse. *J Am Soc Echocardiogr* 2014; 27(6): 601-607.
- 2014 Lonn E, **Rambihar S**, Teo KK et al. Heart rate is associated with increased risk of major cardiovascular events, cardiovascular and all-cause death in patients with stable chronic cardiovascular disease: an analysis of ONTARGET/TRANSCEND. *Clin Res Cardiol* 2014; 103:149–159
- 2010 **Rambihar S**, Dokainish H. Right Ventricular Involvement in Coronary Artery Disease: Focus on non-invasive cardiac imaging. *Curr Opin Cardiol* 2010; 22(5):456-63.
- 2010 Rambihar VS, **Rambihar S**, Rambihar V. Editorial: Heart Disease and South Asians 50 years later: a time for change. *Heart* 2010 Jul; 96 (14): 1168. (Letter to Editor)

- 2010 Rambihar VS, **Rambihar S**, Rambihar V. Race, Ethnicity and Heart Disease: A Challenge for Cardiology for the 21st Century. Am Heart J 2010; 159(1): 1-3. (Editorial)
- 2007 Salehian O, Schwerzmann M, **Rambihar S**, Liu P. Left Ventricular Dysfunction and Mortality in Adult Patients with Eisenmenger Syndrome. Congenit Heart Dis 2007;2:156–164.
- 2007 **Rambihar S**, Abramson B. Cardiovascular Imaging and Noninvasive Diagnosis for Older Adults. Geriatrics and Aging 2007;10(1): 14-22.
- 2006 Therrien J, **Rambihar S**, Granton J. Eisenmenger syndrome and atrial septal defect: nature or nurture? Can J Cardiol 2006 Nov;22(13):1133-6.

PROFESSIONAL MEMBERSHIP:

Member, Canadian Society of Echocardiography
Member, American Society of Echocardiography, Membership #120234
Member, Canadian Cardiovascular Society, Membership #129991
Member, American College of Cardiology, Membership #897110
Member, Ontario Medical Association, Membership #0837682
Member, Canadian Medical Association, Membership #12367