Endovascular Simulation: A Systematic Review of the Validity Evidence, Methodology, Quality and Outcomes

.

BY

MARK G. DAVIES
B.A., University of Dublin, Dublin, 1986
M.B., B.Ch., B.A.O., University of Dublin, Dublin, 1986
Ph.D., University of Dublin, Dublin, 1996
M.D., University of Dublin, Dublin, 1999
M.B.A., University of Rochester, Rochester, 2008

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, 2019

Chicago, Illinois

Defense Committee:
Yoon Soo Park, Ph.D., Chair and Advisor
Ara Tekian, PhD
Dorthea Juul, Ph.D., American Board of Psychiatry and Neurology

Table of Contents

List of Figures

List of Tables

Abbreviations

ABMS	American Board of Medical Specialties
ABS	American Board of Surgery
ACGME	Accreditation Council for Graduate Medical Education
MERSQI	Medical Education Research Study Quality Instrument
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies, version 2
RRC	Residency Review Committee for Surgery
GRRAS	Guidelines for reporting reliability and agreement studies
STARD	Standards for Reporting of Diagnostic Accuracy

Summary

Vascular surgery as a specialty has undergone both a procedural and a training transformation over the last 10 years with a greater emphasis has been placed on endovascular simulation to accelerate skill acquisition at each level of training. The aim of this study is to evaluate the validity evidence, research methods, reporting quality and outcomes that support simulation training in endovascular interventions. An electronic search for relevant articles published between January 2000 and December 2018 was performed and identified sixty-six reports met the inclusion criteria. The skill sets assessed in these studies were basic skill sets in 24%, moderately difficult skills in 14%, complex skill set in 62%. When one examines the prevalence of key educational features of simulation, clinical variation, repetitive practice and feedback were the dominant features employed. Applying the Best Practice for Simulation to the 66 studies chosen. The average number of elements (max 11) identified was 5□3 (mean□SD) or under 50% of elements for a successful simulation design. On MERSQI, sixty-six percent of the studies achieved a passing score of 12 or greater and were considered adequate. When the STARD/GRRAS criteria for methodology are applied to the studies, the overall methodology was poor with 42% of necessary components (20 out of 47) being accounted for in the studies. Within the QUADAS-2 criteria, there was a bias in the selection of the participants (66% of studies). No studies used either Messick's or Kane's framework of validity. When analyzed using the framework of Messick, all failed to capture all sources of validity evidence. Most studies referenced some validity for content. However, few demonstrated evidence for response process, internal structure, relation to other variables and consequences. When analyzed using the Kane's framework, most of studies reported well on Scoring, however all had weak rationales and discussion of Generalization, Extrapolation and Implications. The research methods and reporting quality for simulation in vascular surgery is weak and requires significant refinement. When two contemporary frameworks of validity are applied, the current body of work fails to achieve sufficient rigor to be valid and further work must be done to strengthen this area of assessment before widespread introduction into Graduate Medical Education or professional examinations.

Chapter 1 - Introduction

As a specialty, vascular surgery continues to undergo both procedural and training transformations (Mills, 2008). Firstly, the increased penetrance of endovascular techniques and other significant innovations in technology have increased the volume of primary endovascular procedures, such that more than 50% of procedures are currently conducted with endovascular techniques (Choi et al., 2001; Goodney, Beck, Nagle, Welch, & Zwolak, 2009). This places a need on training programs to educate and ensure mastery of both the trainers and trainees in the principles of endovascular skills and in each new technology. Secondly, the introduction of integrated vascular residency programs, which take medical students directly into the specialty rather than taking residents, who have completed a five-year general surgery residency and enter as a fellow with a 5-year core surgical skill set, has created a void in competency of basic endovascular and open surgical skills (Lee et al., 2010; Schanzer et al., 2009). Consequently, training programs must provide endovascular training throughout the years of training to ensure that the integrated residents perform at a commensurate level to vascular fellows. Thirdly, the introduction of duty hour restrictions on residents and fellows has impacted the time available for teaching and training and has reduced the scope and volume of the patient material, to which trainees are exposed (Lewis & Klingensmith, 2012). This can reduce experience in both simple and advanced endovascular cases. The absence of a large volume of patient material requires that alternative modalities be employed to get trainees up to speed on basic and advanced endovascular techniques. As a result of these three synergistic and disruptive forces impacting the training environment, the need to provide endovascular procedural education for trainees has increased (Mitchell et al., 2014). Many of the instruments used to train and assess endovascular skills are derived from commercial endovascular simulators, have been predominantly driven by industry and have not been the subject of intense scientific rigor. To date, there has been no systematic evaluation of the validity evidence, research methods, reporting quality and outcomes of endovascular simulation undertaken in the literature. The data supporting their use has not been systematically analyzed and a gap exists in our knowledge of its value (Mitchell et al., 2014).

Validity is an important aspect of all assessment tools and is defined as "appropriate interpretation of test results, and a validation study is a process of collecting evidence to support the interpretations of assessment results" (Joint American Educational Research Association, 2014). While the traditional aspects of validity described validity in terms of content, criterion, and construct (Cronbach & Meehl, 1955; Downing, 2003), the field has migrated to two unified conceptual frameworks - Messick's Framework (Cronbach & Meehl, 1955; Downing, 2003; Messick, 1995) and Kane's Framework (Kane, 2006). This migration is driven because the "*traditional aspects of validity is considered fragmented and incomplete and fail to take into account evidence of the value implications of score meaning as a basis for action and of the social consequences of score use"* (Messick, 1995).

In a recent review of simulation in vascular surgery, Mitchell et al (Mitchell et al., 2014) identified 48 reports: 29 studies examined open vascular skills; 19 examined endovascular skills; 6 examined non-technical skills; and one studied teamwork skills. Most of studies (84%) were conducted within a simulated training environment, four (8%) were conducted in a procedural area, and the remainder (3) were performed in both simulated and real world procedural environments. Checklists and global rating scales were the most commonly used metrics for objective technical skills assessment. These tools were shown to have high inter-rater reliability, construct validity and positive user satisfaction and acceptability. None applied a modern framework of validity as described by Messick or Kane. In general, Mitchell et el (Mitchell et al., 2014) concluded that these tools were considered not very practical as they are either procedure-specific or too long (checklist of up to 62 items), making the assessment process labor-intensive, time-consuming, costly, and impractical in evaluation of varying procedures. A second systematic review by See et al (See, Chui, Chan, Wong, & Chan, 2016) reported that contemporary evidence shows that performance metrics within endovascular simulations improve with simulation training. Successful translation to *in vivo* situations is observed in patient specific procedure rehearsals. However, there is no evidence to show that simulation can definitively improve patient outcomes (predictive validity), which is the ultimate desired consequence. A recent expert panel of the Society for Cardiovascular Angiography and Intervention's (SCAI) Simulation Committee concluded that at the present time, simulation lacks a large body of evidence for its use, but they did not comment on the quality of the research available on simulation (Green et al., 2014).

Cook et al (Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014) performed a systematic literature search of original research that evaluated the validity of simulation-based assessment scores using two or more evidence sources. Among 217 eligible studies identified by the authors, only six (3%) referenced Messick's five-source framework, and 51 (24%) made no reference to any validity framework (Cook et al., 2014). The most common concepts mentioned were relationship to other variables (94% of studies reported variation in simulator scores across training levels), internal structure (76% supported reliability data or item analysis), and content (63 %; reported expert panels or modification of existing instruments). Evidence of response process and consequences were each found in <10% of studies (Cook et al., 2014). Cook et al (Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013) also commented in a second paper on simulation-based assessment in health professions that the methodological and reporting quality of assessment studies leaves much room for improvement and could lead to biased results. Borgersen (Borgersen et al., 2018) systematically reviewed the simulation literature and identified 498 studies with a total of 18,312 participants between 2008 and 2017. The authors found that only 6.6% of the studies used a recommended contemporary framework of validity (Messick) and that majority of studies used outdated frameworks such as face validity.

The aim of this study is to perform a systematic review of the current literature on endovascular simulation and evaluate the validity evidence using two modern frameworks (Messick and Kane), examine the quality of research methods and reporting and discuss the outcomes of the studies.

Chapter 2 - Methods

Study Design

An electronic search for relevant articles listed between January 2000 and December 2018 was performed to identify reports and publications on endovascular simulation. The search adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards for systematic reviews (Moher, Liberati, Tetzlaff, & Altman, 2009). Selected studies were then reviewed, assessed for research quality, research bias, graded for evidence of validity using the accepted frameworks of Messick (Messick, 1995) and of Kane (Kane, 2006).

Conceptual Frameworks of Validity

Validity is defined as "appropriate interpretation of test results, and a validation study is a process of collecting evidence to support the interpretations of assessment results" (Joint American Educational Research Association, 2014). The traditional framework of validity that described validity in terms of content, criterion, and construct (Cronbach & Meehl, 1955; Downing, 2003) has migrated to two unified conceptual frameworks - Messick's Framework (Cronbach & Meehl, 1955; Downing, 2003; Messick, 1995) and - Kane's Framework (Kane, 2006):

Messick's Framework*:* Messick integrated the three traditional aspects of validity (content, construct and criterion) into a single concept termed construct validity which incoproated the idea of consequential validity (i.e. the effect the assessment has on education) (**Table I**) (Messick, 1995). In his framework, Messick (Messick, 1995) describes five aspects of the single concept of construct validity: "*content, response process, internal structure, relation to other variables and consequences"*. *Content* describes the "relationship between a test's content and the construct it is intended to measure". *Response process* describes the "analyses of responses (actions, strategies, thought processes) of individual respondents or observers. Differences in response processes may reveal sources of variance irrelevant to the construct being measured. It includes instrument security, scoring, and reporting of results”. Internal structure describes the “degree to which individual items within an instrument fit the underlying constructs. It is often reported by measures of internal consistency reliability and factor analysis internal structure”. *Relations to other variables* describe the “relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent/predictive) or negative (divergent/discriminant) depending on the constructs being measured” . *Consequences* are considered to “assessments that are intended to have some desired effect or may have unintended effects”.

Table I MESSICK’S FRAMEWORK FOR VALIDITY

Evidence Source	Definition
Content	The “relationship between a test’s content and the construct it is intended to measure.”
Process response	Analyses of responses (actions, strategies, thought processes) of individual respondents or observers. Differences in response processes may reveal sources of variance irrelevant to the construct being measured, instrument security, scoring, and reporting of results.
Internal structure	Degree to which individual items within an instrument fit the underlying constructs. It is often reported by measures of internal consistency reliability and factor analysis.
Relations to other variables	Relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent/predictive) or negative (divergent/discriminant) depending on the constructs being measured.
Consequences	Assessments are intended to have some desired effect or may have unintended effects

Kane's Framework: In contrast to Messick, Kane (Kane, 2006) has proposed a validity interpretive argument that consists of three elements: the proposed interpretation, the link between observed performance and interpretation, and the link between interpretations and the decision or use to which these interpretations will be applied. In his validity argument, Kane (Kane, 2006) identified four inferences to inform these three elements: "*Scoring* (translating an observation into one or more scores); *Generalization* (using the scoring data as a reflection of performance in a test setting); *Extrapolation* (using the scoring data as a reflection of real-world performance), and *Implications* (applying the scoring data to inform a decision or action)" (**Table II**) (Cook, Brydges, Ginsburg, & Hatala, 2015; Tavares et al., 2018).

Table II KANE'S FRAMEWORK FOR VALIDITY

Evidence Source	**Definition**
Scoring	Refers to the process of moving from an observed performance to an observed score; it includes the scoring rules, rubric and scoring procedures Scoring (translating an observation into one or more scores);
Generalization	Refers to the degree to which the assessment protocol represents all of the theoretically possible clinical events and the degree to which the scoring data as a reflection of performance in a test setting
Extrapolation	Refers to using the scoring data as evidence for how well candidates will perform in future and novel clinical contexts in the real-world performance
Implications	Refers to the process of applying the scoring data to inform a decision or action)

Systematic Review

Literature search strategy: A two level search strategy of the literature was performed. A search for studies within the MEDLINE, EMBASE, PubMed, Web of Science, PsycINFO, The Cochrane Library and Google Scholar electronic databases for reports on this subject was conducted between January 2000 and December 2018. The year 2000 was chosen to coincide with the introduction of the ACGME competencies. The searches were limited to the English language and, in the case of Google Scholar, to the first twenty chronologically listed pages (~200 hits). Search terms are shown in **Table III**. Selected studies were then abstracted, reviewed and evaluated by three independent reviewers

Table III SEARCH TERMS

Simulation
Assessment,
Evaluation,
Catheter-based,
Endovascular
Skill

Study selection: All papers and reports in the English language that examined endovascular simulation were assessed. Reports where the population of interest includes undergraduate and post graduate surgical trainees and faculty and where catheter-based simulation in endovascular interventions was the basis of the report were reviewed. Types of intervention included any task trainer or bench top model that can be used as simulation for catheter-based endovascular training, however, all computer-based simulators including virtual reality and software imaging were excluded as they did not test the manual skill domains. Reviews and commentaries were also excluded.

Process of conducting the ratings: The initial assessment of each paper was performed by the primary author. Thereafter two single-blinded authors (vascular research fellows) were trained to use each of the scoring tools on a set of pilot papers to assure appropriate application and consensus of interpretation of the elements described before they independently completed a full assessment of the remaining papers. Inter-rater reliability of ratings was quantified by calculating intra-class correlation coefficients for each element of the scoring tools.

Data Synthesis

Description of Studies. Each report was assessed for number of participants, specialty and level of training was recorded for each simulation the study design, skill set employed, and metrics used were assessed. Skill sets ranged from basic to complex. Simulation educational strategies and educational outcomes as previously described by Cook (Cook et al., 2011; Cook et al., 2008) were also examined. Finally, we applied the standards of best practice in simulation design to each of the reports. The International Nursing Association for Clinical Simulation and Learning Standards of Best Practice for simulation described 11 key elements of a successful simulation (Lioce et al., 2015). Applying the International Nursing Association for Clinical Simulation and Learning Standards of Best Practice for simulation, the design of the simulations was scored.

Quality of Educational Research: A recent report by the National Academy of Sciences recommended the development of reliable and valid metrics for quantifying the quality of medical education research to promote improved quality in educational journals and enhance the opinion of educational research with federal funding agencies (Towne, Wise, Winters, & Committee on Research in Education - National Research Council, 2004). Reed et al (Reed et al., 2007) developed a10-item toll for assessing the quality of research studies in medical education - Medical Education Research Study Quality Instrument (MERSQI) The MERSQI surveys 6 areas of study quality: "*study design, sampling, type of data (subjective or objective), validity, data analysis, outcomes and originality*" and allows a numerical score for each domain (0-3) which produces a scoring range from 0-18. The MERSQI tool as described by Reed (Reed et al., 2007) was used to assess the quality of the studies **(Table IV)**. As a threshold for high- or low-quality scores, the median of the MERSQI scores was calculated and used (Cook et al., 2011).

Table IV MEDICAL EDUCATION RESEARCH STUDY QUALITY INSTRUMENT

Category	Variable	Weighting
Study design	Single group cross-sectional or single group post-test only	1
	Single group pre-test and post-test	1.5
	Non-randomized, 2 group	2
	Randomized controlled trial	3
Sampling	**No. of institutions studied**	
	1	0.5
	2	1
	>2	1.5
	Response rate, %	
	Not applicable	0
	50 or not reported	0.5
	50-74	1
	≥75	1.5
Type of data	Assessment by study participant	1
	Objective measurement	3
Validity of evaluation instruments	**Internal structure**	
	Not applicable	0
	Not reported	0
	Reported	1
	Content	
	Not applicable	0
	Not reported	0
	Reported	1
	Relationships to other variables	
	Not applicable	0
	Not reported	0

	Reported	1
Data analysis	Appropriateness of analysis	
	Data analysis inappropriate for study design or type of data	0
	Data analysis appropriate for study design and type of data	1
	Complexity of analysis	
	Descriptive analysis only	1
	Beyond descriptive analysis	2
Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts	1
	Knowledge, skills	1.5
	Behaviors	2
	Patient/health care outcome	3
Originality	Original research	1

Quality of methodology and reporting: A recent report by Cook et al (Cook et al., 2013) utilized a framework derived from the Standards for Reporting Diagnostic Accuracy (STARD) and the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) to determine the quality of the methodology in simulation papers **(Table V)** (Bossuyt et al., 2015; Cook et al., 2013; Kottner et al., 2011). As a threshold for high- or low-quality scores, the median of the STARD/ GRRAS scores was calculated and used.

Table V STANDARDS FOR REPORTING DIAGNOSTIC ACCURACY (STARD) AND THE GUIDELINES FOR REPORTING RELIABILITY AND AGREEMENT STUDIES (GRRAS).

Study Component	STARD Item		GRRAS Item		Reporting element (operational definition)
Title/abstract					
	Identified as study of simulation		Identified that reliability was investigated		Title or abstract identifying the study as an evaluation of the validity, reliability, or diagnostic accuracy of an assessment tool diagnostic accuracy of an assessment tool
					Title or abstract identifying the study as focused on assessment, but not as a study of validity, reliability, or diagnostic accuracy
Introduction					
	Research question				Explicit question, purpose, or hypothesis
					Proposed validity argument (strategy for interpreting validity evidence to be presented)
			Instrument name, description		Description of index test task
			Existing evidence for this simulator		Critical review of evidence relevant to assessment of that construct
Method					
	Study population: eligibility, setting		Study population		Trainee population (eligibility criteria)
					Setting (educational [e.g., simulation laboratory versus clinical)
	Participant recruitment				Identification of eligible trainees (any method defined)
	Participant sampling: consecutive, random		Sampling method		Sampling strategy (any method defined)
	Data collection: prospective or retrospective				Prospective or retrospective data collection

	Reference standard: definition, rationale				Rationale for relationship between index and reference test
	Index test and reference standard: methods		Measurement procedures described		Methods/procedures for index test
	Index test and reference standard: classification				Passing standard
	Raters: number, training		Rater population		Rater population (eligibility cri
					Rater training (done or not done
			Rater number		Rater total number
			Rater characteristics		Rater specialty
	Raters: blinded		Raters independent		Raters blinded to trainee (done or not done)
					Raters blinded to other raters (done or not done)
					Raters blinded to results of reference test (done or not done)
	Statistical methods—accuracy: defined		Statistical methods		All statistical methods defined: comparisons among groups or correlation
	Statistical methods—reliability: defined				All statistical methods defined: reliability
			Sample size calculations (planned)		Sample size calculations
Results					
	Study dates				Study dates
	Participant demographics		Participant number		Trainee number enrolled
			Characteristics, participants		Trainee training level
	Participants eligible, not enrolled; flow diagram				Trainee number eligible
					Flow diagram

	Time and events between index and reference tests				Time interval between index and reference test
	Severity of disease in population				Trainee baseline proficiency: objective measurements
					Trainee baseline proficiency: prior experience with that task
	Distribution of test results				Central tendency (mean, median) and variability (standard deviation, range) for scores
					Central tendency (mean, median) without variability
					Figure (scatter plot) or table (contingency table)
	Adverse events				Consequences of testing, adverse or beneficial
	Estimates of accuracy and statistical				Estimate of accuracy (correlation coefficient or other)
					Receiver operating characteristic (ROC) curve, sensitivity, or specificity of test
					Confidence intervals for accuracy estimates
	Indeterminate and outlier results: how handled				Scoring process described
					Indeterminate and outlier results considered in scoring
	Variability across subgroups of participants or raters				Subgroup analyses interpreted as relating to score validity
	Reproducibility		Reliability, including uncertainty		Reliability (any)
					Confidence intervals for reliability estimates
Discussion					
	Clinical applicability		Practical relevance		Confidence intervals for reliability estimates of validity evidence)

Quality of unbiased results: The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) is a seven-question tool, which covers four domains of participant selection, index test, and reference test and is designed to determine possible bias (systematic flaws that distort study results), applicability to the study question and evaluates bias in study flow **(Table VI)** (Cook et al., 2013; Kottner et al., 2011).

Table VI THE QUALITY ASSESSMENT OF DIAGNOSTIC ACCURACY STUDIES (QUADAS-2)

Domain	Patient Selection	Index Test	Reference Standard	Flow and Timing
Description	Describe methods of participant selection	Describe the index test and how it was conducted and interpreted	Describe the reference standard and how it was conducted and interpreted	Describe any participants who did not receive the index tests or reference standard or who were excluded from the 2x2 table
	Describe included participant (previous testing, presentation, intended use of index test, and setting)			Describe the interval and any interventions between index tests and the reference interventions between index tests and the reference standard
Signaling questions (yes, no, or unclear)	Was a consecutive or random sample of patients enrolled?	Were the index test results interpreted without knowledge of the results of the reference standard?	Is the reference standard likely to correctly classify the target condition?	Was there an appropriate interval between index tests and reference standard?
	Was a case–control design avoided?	If a threshold was used, was it prespecified?	Were the reference standard results interpreted without knowledge of the results of the index test?	Did all participants receive a reference standard?
				Did all participants receive the same reference standard?
	Did the study avoid inappropriate exclusions?			Were all patients included in the analysis?

Risk of bias (high, low, or unclear)	Could the selection of participants have introduced bias?	Could the conduct or interpretation of the index test have introduced bias?	Could the reference standard, its conduct, or its interpretation have introduced bias?	Could the participant flow have introduced bias?
Concerns about applicability (high, low, or unclear	Are there concerns that the included participants do not match the review question?	Are there concerns that the index test, its conduct, or its interpretation differ from the review question?	Are there concerns that the target condition as defined by the reference standard does not match the review question?	

Quality of Validity Evidence: The grading scale to evaluate the quality of Messick's sources of validity evidence derived by Ghaderi et al (Ghaderi et al., 2015) was used: The scale was 4 points (0-3) and was defined as follows: 0 = no discussion or data presented as a source of validity evidence; 1 = data that weakly support the validity of score interpretations; 2 = some data (intermediate level) that support the validity of score interpretations, but with gaps; and 3 = multiple sets of data that strongly and completely support the validity of score interpretations **(Table VII)**.

Table VII CRITERIA FOR RATING VALIDITY EVIDENCE IN A MESSICK'S FRAMEWORK

Content	
0	No discussion or data regarding the instrument content
1	Only discussion or limited amount of data (simply listing items without justification)
2	Listing assessment themes with some references and justifications, limited description of the process for creating the instrument Alternatively, reference to a prior study on an assessment instrument that meets these criteria
3	Well-defined process for developing instrument content, including both an explicit theoretical/conceptual basis for instrument items and systematic item review by experts
Response process	
0	No discussion or data regarding the response process
1	Minimal discussion and limited data presented. Use of an instrument without reporting the results. Discussing the impact of response rate on assessment scores or speculating on the thought processes of learners
2	Some data regarding thought processes and analysis of responses. Some data about implication of systems that reduced response error
3	Multiple sources of supportive data, including critical examination of thought processes, analysis of responses for evidence of halo error or rater leniency, or data demonstrating low response error
Internal structure	
0	No discussion or data
1	Minimal data with regard to internal structure, some reliability with a single measure
2	Factor analysis incompletely confirming anticipated data structure, or a few measures of reliability reported
3	Factor analysis confirming anticipated data structure or multiple measures of reliability. Item analysis data, item/test characteristic curves (ICCs/TCCs), inter item correlations, item-total correlations), generalizability analysis
Relation to other variables	
0	No discussion or data
1	Correlation of assessment scores to outcomes with minimal theoretical importance, a single measure of validity (relationship between level of training and scores)
2	Correlation of assessment scores to outcomes with some theoretical importance

3	Correlation (convergence) or no correlation (divergence) between assessment scores and theoretically predicted outcomes or measures of the same construct. Such evidence will usually be integral to the study design and anticipated a priori, generalizability evidence
Consequences	
0	No discussion or data
1	Limited data about the consequences of the assessment. Merely discussion about the consequences of assessment (e.g., data regarding usefulness of assessment based on post-assessment survey)
2	Description of consequences of assessment that could conceivably impact the validity of score interpretations (although these impacts are not explicitly identified by the authors)
3	Description of consequences of assessment that clearly impact on the validity of score interpretations, as supported by data and convincingly argued by the authors. Evidence will usually be integral to the study design and anticipated a priori Such

A new grading scale for Kane's Validity framework (elements and inferences) was devised based on the principles used by Ghaderi et al (Ghaderi et al., 2015) to develop the grading scale for Messick's Validity framework. The scale was 4 points (0-3) and was defined as follows: 0=no discussion or data presented as a source of validity evidence; 1= data that weakly support the criteria; 2 = some data (intermediate level) that support the validity of Kane's criteria, but with gaps; and 3 = multiple sets of data that strongly and completely support the validity of Kane's criteria, **(Table VIII)**.

Table VIII CRITERIA FOR RATING VALIDITY EVIDENCE IN KANE'S FRAMEWORK

Statement of the proposed interpretation	
0	No statement of the proposed interpretation No discussion or data to establish a rationale for the proposed interpretation
1	Limited discussion of justification for the proposed interpretation
2	Greater than limited discussion of justification for the proposed interpretation
3	Well-defined statement of the proposed interpretation, which includes explicit theoretical/conceptual bases for the proposed interpretation
Linking observed performance to an interpretation	
0	No statement of the proposed observed performance No discussion or data to establish a linkage between observed performance to the defined interpretation
1	Minimal discussion and limited data to establish a linkage between observed performance to the defined interpretation
2	Greater than minimal discussion and limited data to establish a linkage between observed performance to the defined interpretation
3	Multiple sources of supportive data to establish a linkage between the observed performance and the defined interpretation
Linking the interpretation to a decision	
0	No statement of the proposed decision No discussion to establish a linkage between the defined interpretation to a defined decision
1	Minimal data with regard to linkage of the interpretation to a decision
2	Identified sources (<2) of supportive data with regard to linkage of the interpretation to a decision
3	Multiple sources (>2) of supportive data with regard to linkage of the interpretation to a decision

Scoring	
0	No discussion of translating an observation into one or more scores
1	Minimal discussion of translating an observation into one or more scores
2	Greater than minimal discussion of translating an observation into one or more scores
3	Strong rationale for translating an observation into one or more scores
Generalization	
0	No discussion or data regarding generalization
1	Minimal discussion of generalization of one or more scores as a reflection of performance in a test setting
2	Greater than minimal discussion of generalization of one or more scores as a reflection of performance in a test setting
3	Strong rationale for using the scoring data as a reflection of performance in a test setting
Extrapolation	
0	No discussion or data regarding extrapolation of performance to a real-life setting
1	Minimal discussion of extrapolation of performance in a real-life setting
2	Greater than minimal discussion of extrapolation of performance in a real-life setting
3	Performance data reflects real-world performance
Implications	
0	No discussion or data regarding implications
1	Minimal discussion or data regarding implications in practice or career
2	Greater than minimal discussion or data regarding implications in practice or career
3	Applying performance to inform a decision or action in practice or career

Outcomes: Outcomes of the simulation were distinguished using Cook's Modification of Kirkpatrick's classification where the outcomes of learning were categorized by knowledge, skills, behaviors and results (patient effects). The skill category was subclassified as time, process, and product while behaviors with patients was segmented in time and process measures (Cook et al., 2012; Cook et al., 2011). In addition, a catalogue of outcomes from the studies was created and a common set of outcome measures identified.

Statistical Analysis

Statistical Analysis: Measured values are reported as percentages or mean±SD. Inter-rater reliability was quantified by calculating kappa and intraclass correlation coefficients for the individual element of each Instrument employed.

.

Chapter 3 - Results

Literature Search. Following a literature search, 701 articles were identified and 120 were deemed suitable for full review. The PRISMA flow diagram is shown in **Figure 1**. Sixty-six studies were identified that met the criteria for analysis. Forty-two papers were published between 2011 and 2018, while 24 papers were published between 2000 and 2010 (**Figure 2**). Ninety-seven percent of the studies were considered original and the remainder were continuations of previous work but with new elements.

Figure 1 PRISMA Flow Diagram

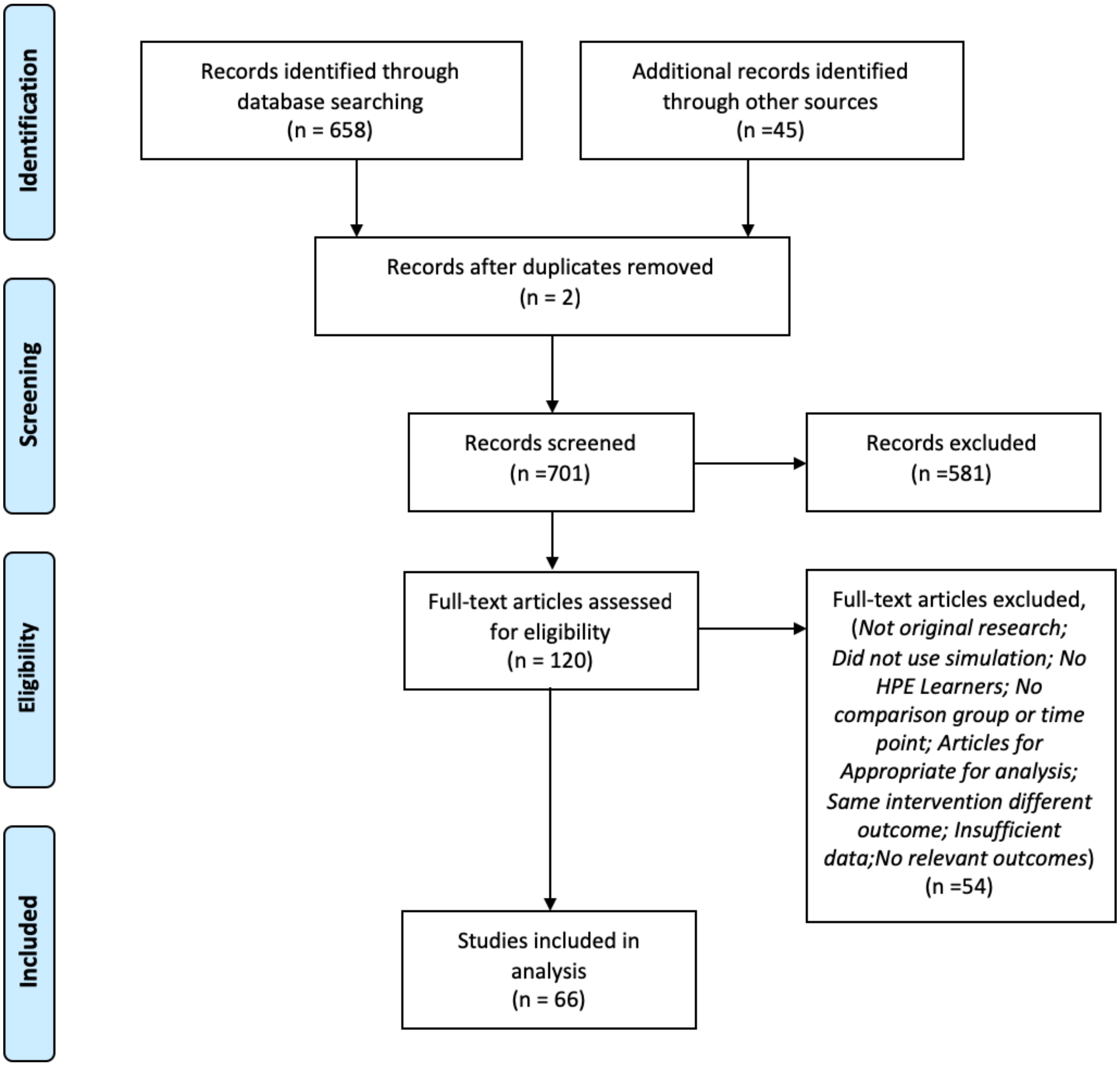
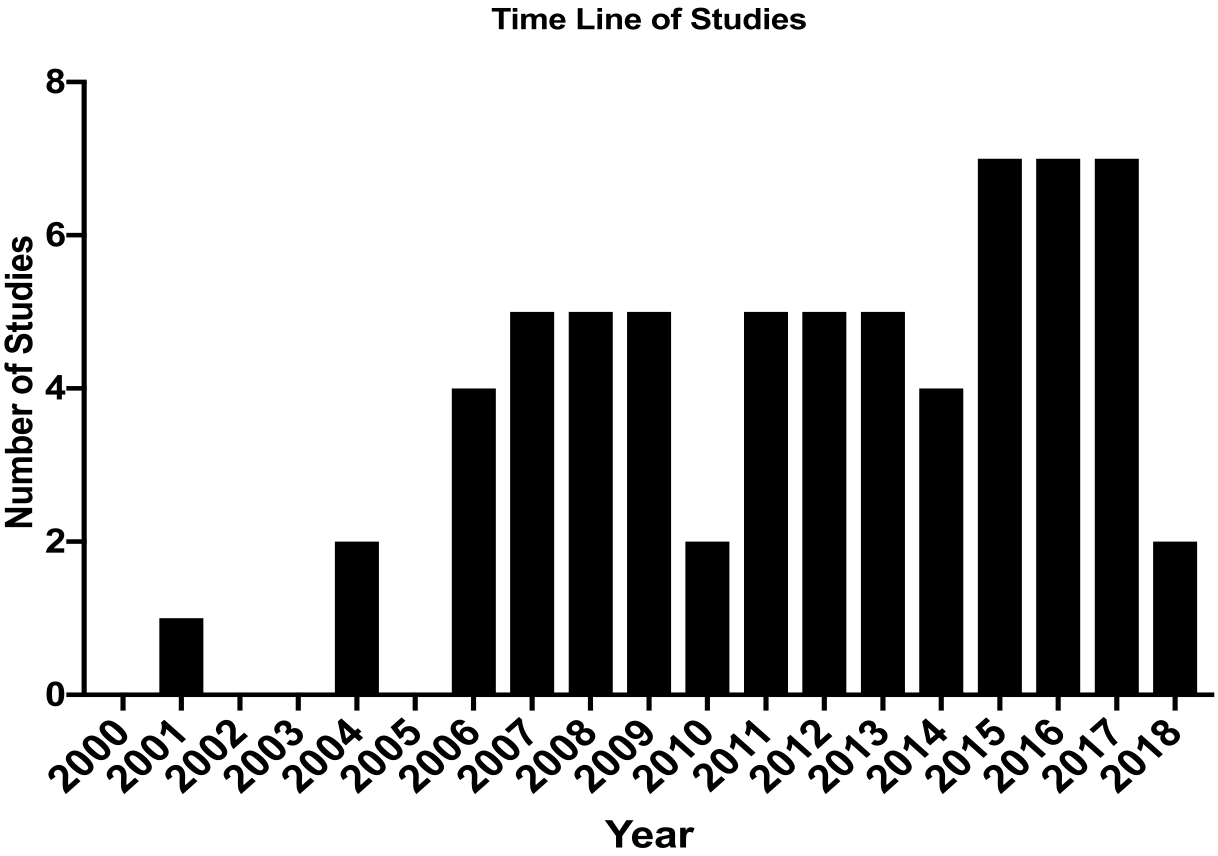

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 658)
Additional records identified through other sources (n =45)
Records after duplicates removed (n = 2)
Records screened (n =701)
Records excluded (n =581)
Full-text articles assessed for eligibility (n = 120)
Full-text articles excluded, (Not original research; Did not use simulation; No HPE Learners; No comparison group or time point; Articles for Appropriate for analysis; Same intervention different outcome; Insufficient data;No relevant outcomes) (n =54)
Studies included in analysis (n = 66)

Figure 2 **Distribution of Publications by year**



Study Characteristics: Of the 66 studies identified, 52% had a study design that consisted of a pre and post interventional comparison, while the remainder were considered observational (**Table IX**). Of those that were a pre- and post- interventional comparison only 30% randomized their participants (**Table IX**). The skill sets tested in these studies were basic catheter skill sets in 24%, moderately difficult (intermediate) catheter skill sets in 14%, complex skill sets in 52% and a mixed composition of catheter skill sets in 10% (**Table IX**).

Table IX STUDY CHARACTERISTICS

Studies	66
Type	
Observational	48%
Pre- and post-intervention comparison	52%
Design	
Non randomized	80%
Randomized	20%
Skill Set	
Basic Skill Set	
Catheter skills	8%
Aorta angioplasty	2%
Iliac angioplasty	9%
REBOA	5%
Moderate Skill Set	
Superficial Femoral Artery Angioplasty (SFA)	2%
Abdominal and thoracic Endovascular Aortic Repair (EVAR)	12%
Complex Skill Set	
Carotid Artery Stenting (CAS)	30%
Renal Artery Stenting (RAS)	11%
Coronary Angioplasty	11%
Mixed skill Sets	
Iliac /Renal Interventions	6%
Iliac/ SFA Interventions	2%
RAS/CAS/EVAR/SFA	2%

Participants were heterogeneous, ranging from medical students, to novice, and from intermediate to skilled interventionalists. The total number of participants was 1453 and the average number of participants in the studies was 22 (SD 11, range 4-77). The participants were drawn from multiple specialties with vascular surgery being the dominant specialty (**Table X**). The remainder were drawn in descending order from cardiology, Interventional Radiology, Neurovascular Interventionalists, Trauma Surgery and multidisciplinary teams (**Table X**). Forty of the 66 studies used residents as the principal participants, 30 included attending staff, 22 recruited fellows and 15 recruited medical students and/or allied health staff (**Table X**).

Table X PARTICIPANTS

Participants	1453	
Specialty	**Number of studies**	**% of Studies**
Vascular Surgery	41	62%
Cardiology	9	12%
interventional Radiology	6	11%
Neurovascular	6	9%
Trauma Surgery	3	5%
Multidisciplinary	1	2%
Level of Training	**Number of studies**	
Attending staff	30	
Fellows	22	
Residents	40	
Medical Students	11	
Allied Health Staff	4	

When one examines the prevalence of key simulation features, clinical variation, repetitive practice and feedback were the dominant features in most studies (**Table XI**). Additional features were identified based across the studies that could be interpretative as strategies for Individualized Learning (high), Range of Task difficulty (present), Multiple Learning Strategies (high), Distributive Practice (>1 day), Blended Learning, Mastery Learning (Present) and Cognitive Interactivity (high) (**Table XI**). None mentioned Deliberate Practice (**Table XI**).

Table XI SIMULATION KEY FEATURES

	Number of studies	**% of Studies**
Clinical Variation (present)	49	74%
Repetitive Practice (present)	26	39%
Feedback (High)	24	36%
Individualized Learning (high)	17	26%
Range of Task difficulty (present)	14	21%
Multiple Learning Strategies (high)	12	18%
Distributive Practice (>1 day)	11	17%
Blended Learning (present)	4	6%
Mastery Learning (present)	4	6%
Cognitive Interactivity (high)	1	2%
Deliberate Practice (present)	0	0%

To quantify the changes in performance in response to these strategies, the various authors used a variety of metrics, which were drawn from machine derived metrics and various rating scales. The majority of studies used time to complete a task as a marker of improved performance which are derived from the particular simulator used in the study. To support this time-based assessment many used a Generic Performance Scale, Procedure Specific Scale or Global Rating Scale (**Table XII**). Most studies used 1-2 raters who were not blinded to the participant, did not receive training in the scales and no inter-rater variation was reported. The performance scales were anchored with expert performance in less than 10% of the studies.

Table XII METRICS

Machine Derived Metrics - Times	43	65%
Machine Derived Metrics - Qualitative	16	24%
Computerized tracking of catheter and haptics	4	6%
Generic Performance Scale	18	27%
Procedure Specific Scale	20	30%
Global Rating Scale	13	20%
Psychometrics	4	6%

The International Nursing Association for Clinical Simulation and Learning Standards of Best Practice for simulation described 11 key elements of a successful simulation design. Applying these standards for simulation to the 66 studies chosen. we found that few performed a needs assessment or a pilot prior to engaging in the simulation event (**Table XIII**). Very few provided an opportunity for a post event evaluation (**Table XIII**). Most of the studies failed to offer a pre-briefing and debriefing. (**Table XIII**). The various reports scored higher on designing a scenario, on providing a facilitative approach, on defining measurable objectives, and on ensuring fidelity (**Table XIII**). The average number of elements identified was 5±3 (mean±SD, max score 11) with a median of 5 and a range of 0-10 elements. Fifty-three percent of the studies scored at the median or higher.

Table XIII STANDARDS FOR SIMULATION

	Criteria	% studies satisfying the criterium	ICC
1	Perform a needs assessment to provide the foundational evidence of the need for a well-designed simulation-based experience.-	3%	1.00
2	Construct measurable objectives.	79%	0.75
3	Structure the format of a simulation based on the purpose, theory, and modality for the simulation-based experience.	76%	0.80
4	Design a scenario or case to provide the context for the simulation-based experience.	69%	0.85
5	Use various types of fidelity to create the required perception of realism.	91%	0.70
6	Maintain a facilitative approach that is participant centered and driven by the objectives, participant's knowledge or level of experience, and the expected outcomes.	55%	0.75
7	Begin simulation-based experiences with a pre-briefing	34%	0.90
8	Follow simulation-based experiences with a debriefing and/or feedback session.	27%	0.90
9	Include an evaluation of the participant(s), facilitator(s), the simulation-based experience, the facility, and the support team.	31%	0.74
10	Provide preparation materials and resources to promote participants' ability to meet identified objectives and achieve expected outcomes of the simulation-based	28%	0.80
11	Pilot test simulation-based experiences before full implementation.	6%	1.00

Study Quality. The conduct of the studies was assessed using three metric systems: overall quality of the studies was examined using MERSQI, study methodology was assessed by STRAD/GRRAS criteria and bias was quantified by QUADAS-2 criteria.

The median MERSQI score was 12, and this was chosen as the threshold for an adequate study in the field (Cook et al., 2011). Sixty-six percent of the studies achieved a MERSQI score of 12 or greater and were considered adequate (**Figure 3** and **Table XIV**). The majority of studies were single institution, single group cross-sectional or single group post-test only studies (**Appendix Table AI**). While most used objective measurement, the validity evidence supporting the instruments was weak (**Appendix Table AI**). Data analysis was appropriate with most concentrating on descriptive statistics.

Figure 3 **Distribution of MERSQI Scores**

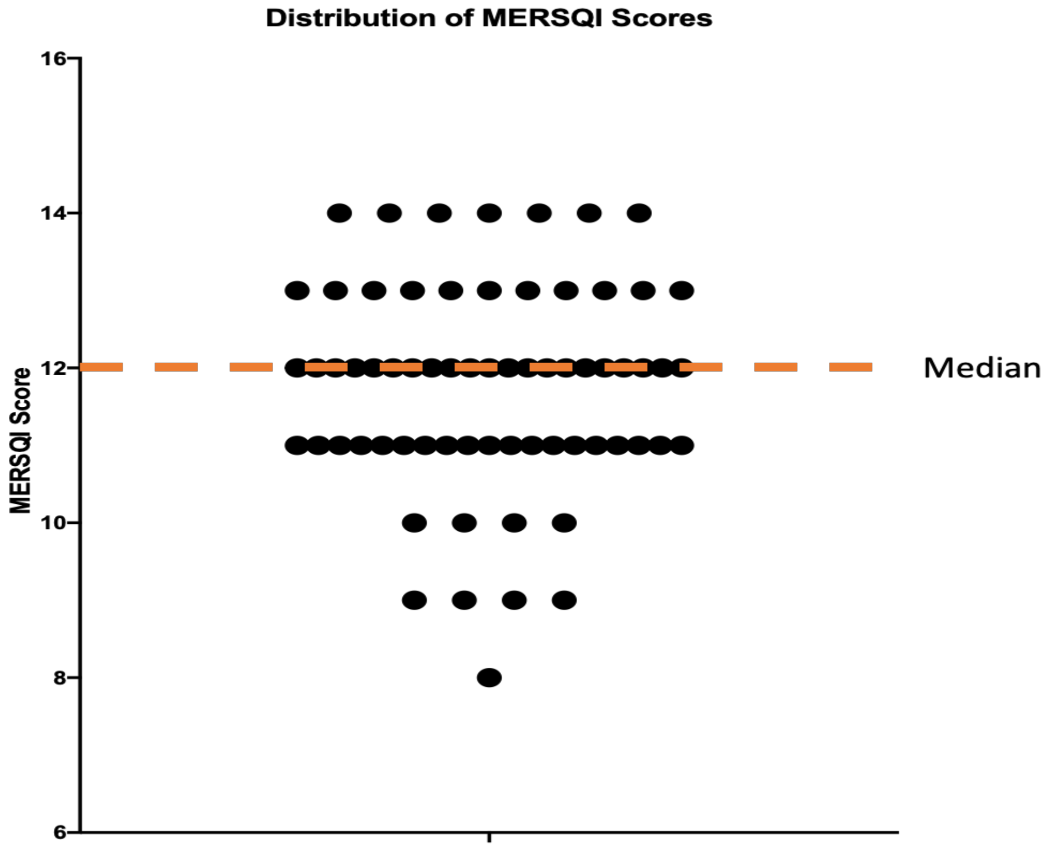


Table XIV REPORTING QUALITY AS DETERMINED BY MERSQI

	Number of Criteria	Average number of components present	SD	Median % component present	ICC
Study design	3	1.6	0.1	52%	0.80
Sampling	3	1.7	0.1	57%	0.80
Type of data	3	3.0	0.2	100%	0.80
Validity of evaluation instruments	3	1.5	0.0	50%	0.84
Data analysis	3	2.5	0.1	83%	0.71
Outcomes	3	1.5	0.1	49%	0.93
Originality	1	1.0	0.1	97%	0.93

	Criteria			
	0	1	2	3
Study design		0%	89%	11%
Sampling		0%	94%	6%
Type of data		0%	8%	92%
Validity of evaluation instruments		0%	86%	14%
Data analysis		0%	29%	71%
Outcomes		0%	97%	3%
Originality	3%	97%		

When the STARD/GRRAS criteria for methodology are applied to the studies, the overall methodology was poor with 42% of necessary components (20 out of 47) being accounted for in the studies. The distribution of STARD/GRRAS scores is shown in **Figure 4**. The presence of the necessary components in each section of the paper were abstract - 52%. introduction - 56%, methods - 46%, results - 38%, and discussion - 63% (**Table XV**). The best described component in the introduction was the description of the index test while the poorest description was discussion of the proposed validity argument (**Appendix Table AII**). In the methods, descriptions of the data collection were well done, while sampling strategy was mentioned in only 16% of the reports (**Appendix Table AII**).   Descriptions of the rating procedures, rater selection and rater training were poor across the studies (**Appendix Table AII**). Only 55% of the reported studies stated a passing standard. In the results, few studies identified the dates of study and few showed a flow diagram (**Appendix Table AII)**. Analysis of outcomes and potential consequences was basic and sophisticated statistics (**Appendix Table AII**). In the discussion sections of the papers, clinical or practical relevance was discussed in the majority of the papers (**Appendix Table AII**).

Figure 4 **Distribution of STARD/GRRAS Scores**

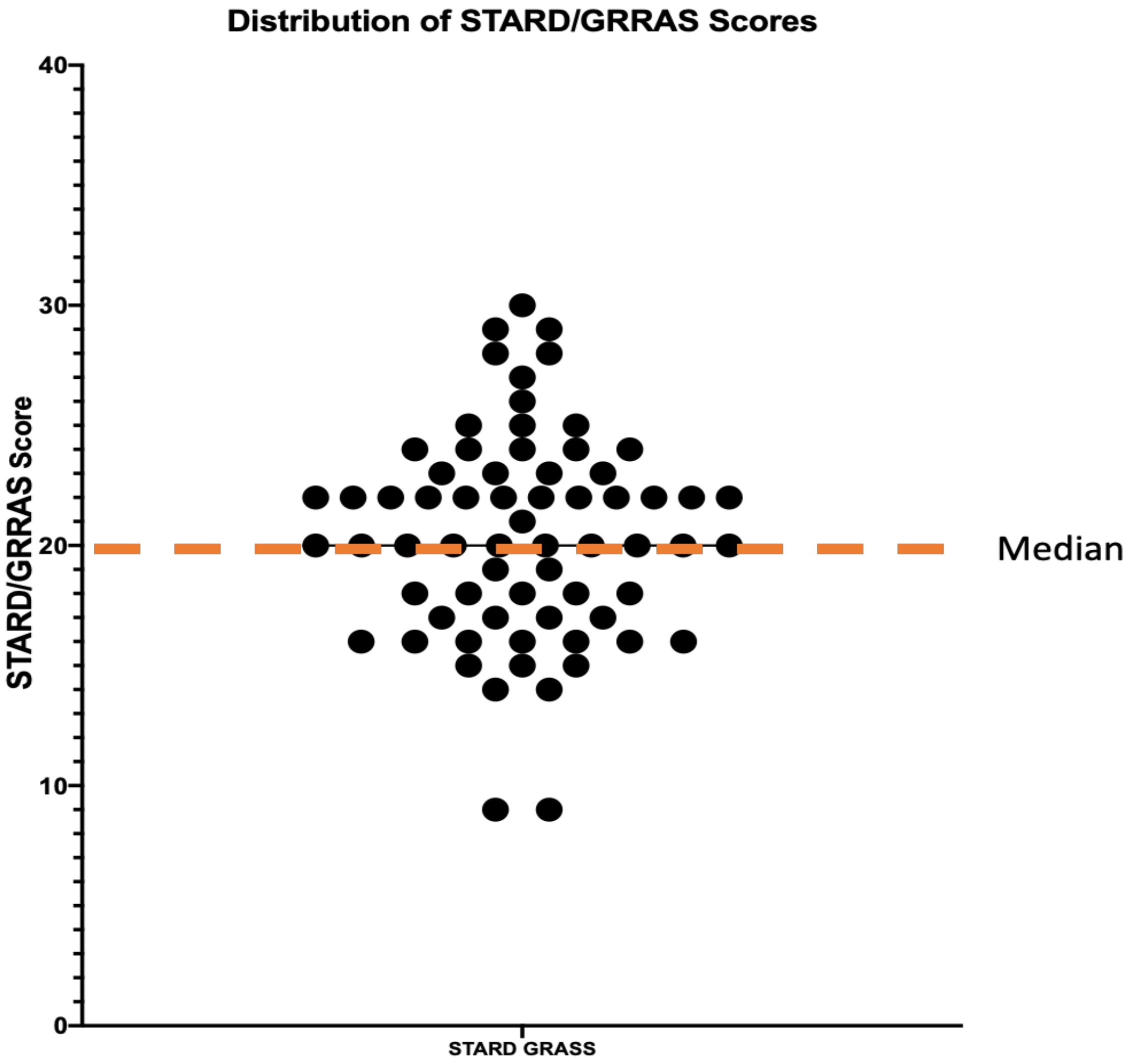


Table XV REPORTING QUALITY AS DETERMINED BY COMPOSITE STARD / GRRAS CRITERIA

	Number of Criteria	**Median % component present**	**Average number of components present**	**SD**	ICC
Title/abstract	1	10%	1	0	1.00
Introduction	4	50%	2.24	0.75	0.88
Methods	20	44%	8.38	2.60	0.87
Results	20	35%	7.18	2.20	0.86
Discussion	1	100%	0.83	0.39	0.78

When the studies were evaluated using the QUADAS-2 criteria, the process of participant selection scored well across the 5 criteria in that section (**Appendix Table AIII**). The distribution of the component scores, concern for applicability and concern for bias are shown in **Figures 5 and 6**. The processes associated with the index and reference tests were handled well and most studies satisfied the stated criteria. (**Appendix Table AIII**). Flow and Timing of the study was also acceptable (**Appendix Table AIII**). Within the QUADAS-2 criteria, there was a bias in the selection of the participants (66% of studies). Application of the index and reference tests also raised the concern for bias in up to half of the studies. The flow of participants was considered to have a low bias (**Table XVI**). Application of the index and references tests was appropriate in 82% of the studies and the flow of participants was also considered applicable (**Table XVI**).

Figure 5 Distribution of QUADAS-2 Scores

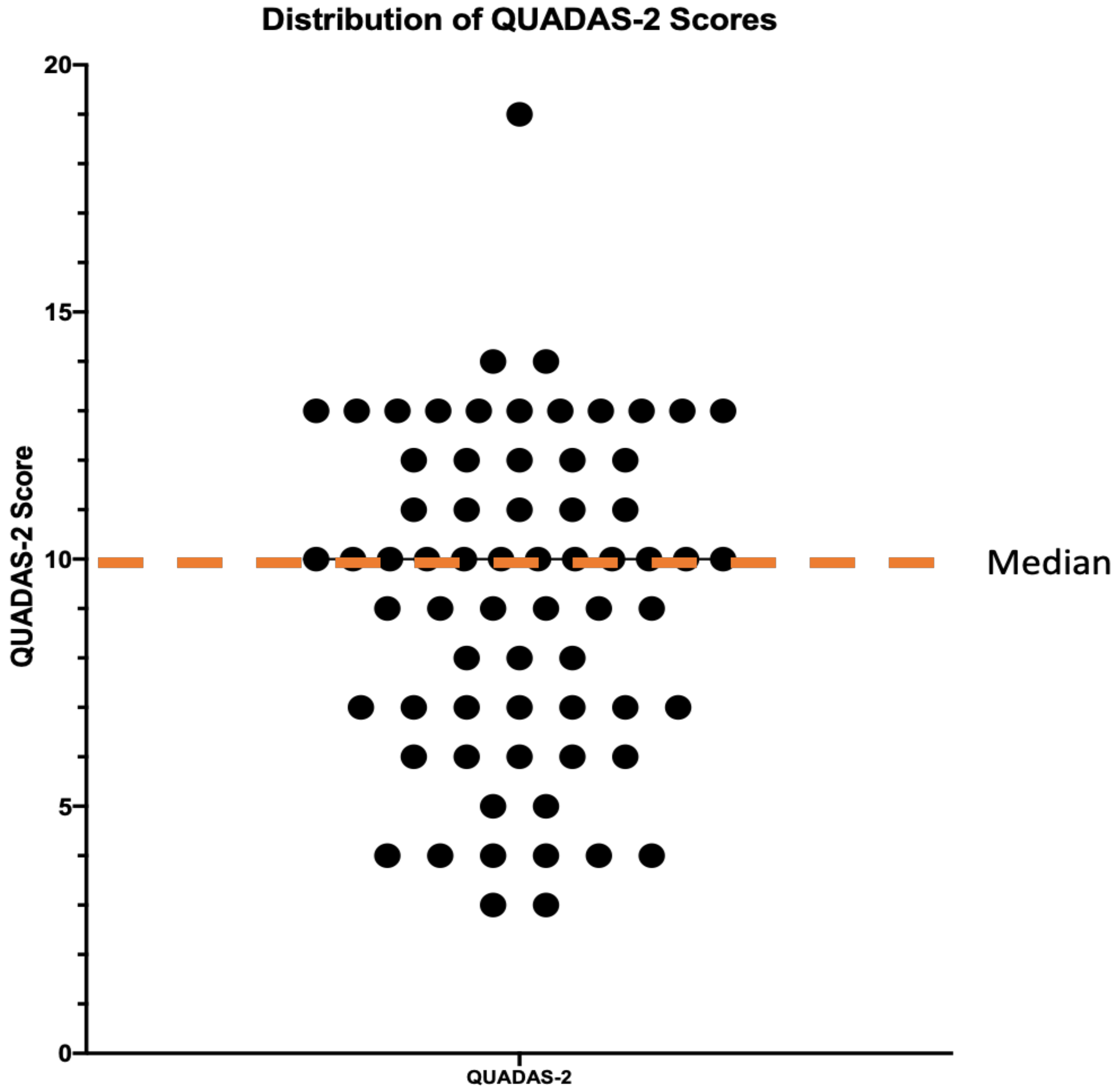


Figure 6 Distribution of QUADAS-2 Bias and Appiicability Determinations

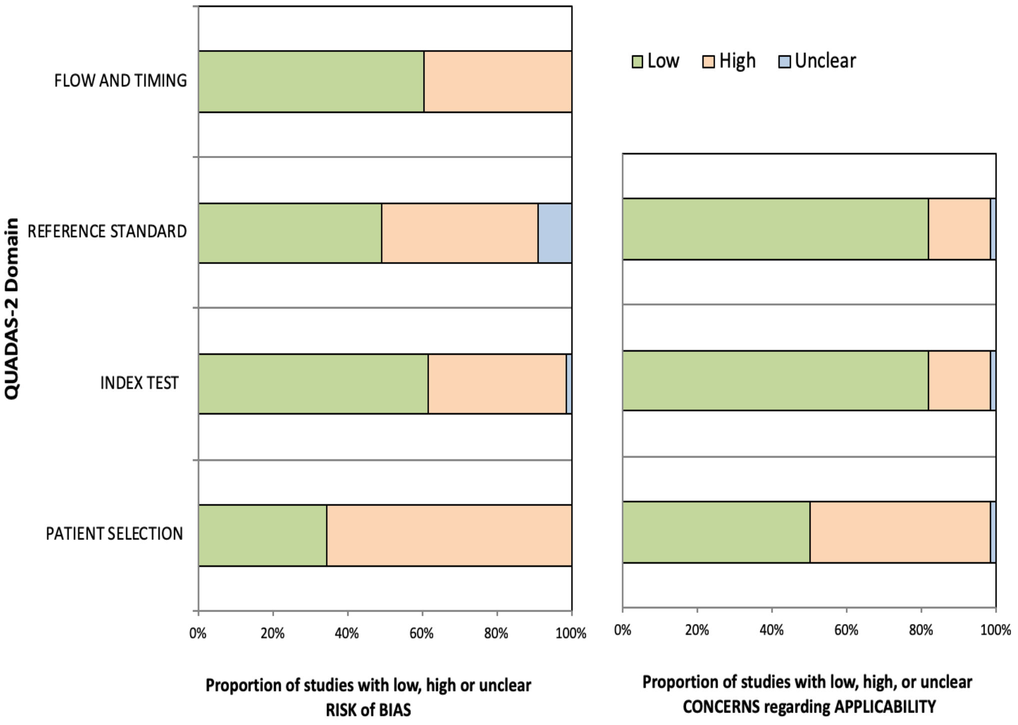


Table XVI QUALITY ASSESSMENT OF BIAS (QUADAS-2)

Criteria	**Parameter**		ICC
Participant selection	Low risk of bias	34%	0.70
	Low concern about applicability	51%	0.84
Index test: conduct or interpretation	Low risk of bias	63%	0.60
Index test: match with target condition	Low concern about applicability	82%	0.68
Reference test: conduct or interpretation	Low risk of bias	49%	0.64
Reference test: match with target condition	Low concern about applicability	82%	0.60
Flow of participants	Low risk of bias	60%	0.86

Validity:

When each of the papers were reviewed, there was limited emphasis on validity across the papers reviewed. Construct validity was mentioned in 28%, face validity in 22% and content validity in 8%. Most relied on an assumption that the use of the simulator and a rating instrument conveyed validity to the study. With little legacy inferences to validity present, we sought to examine the presence of sufficient data to support either Messick's framework or Kane's framework of validity.

Messick's Framework: Messick's framework consolidates prior assessments of validity into a five-category unified framework. When the 66 studies were analyzed using the proposed grading scale for Messick's framework (**Table VII**), most studies reported moderate validity evidence for content (mean score 2); however, few demonstrated good evidence (score of 2 or greater) for response process (mean score 1), internal structure (mean score 1), relations to other variables (mean score 1) and consequences (mean score 1) (**Table XVII**). The average score for Messick framework using the current grading scale was 6±2 (mean±SD) with a median of 6 (interquartile 25th and 75th- 2 and 7.5; range 2-13; scoring range 0-15),

Table XVII MESSICK'S FRAMEWORK

Evidence Source							
	Score	Median	Min	Max	Average	SD	ICC
Content	0-3	2	0	3	2	1	0.80
Process Response	0-3	2	0	3	1	1	0.84
Internal structure	0-3	1	0	3	1	1	0.86
Relations to other variables	0-3	1	0	3	1	1	0.68
Consequences	0-3	1	0	3	1	1	0.74
Overall Score	15	6	2	13	6	3	

Kane's Framework: Kane's framework offers a simpler and refined three category framework derived from four inferences for examining the validity of a study. Using the new grading scale proposed in this paper for Kane's framework (**Table VIII**), most of studies reported a statement of the proposed interpretation (mean score 2); and provided good evidence of Linkage of observed performance to an interpretation (mean score 2); in contrast, there was weak evidence of Linkage of the interpretation to a decision (mean score 1) (**Table XVIII**). The average score for Kane's framework using the current grading scale was 5±1 (mean±SD) with a median of 5 (interquartile 25th and 75th- 4 and 5; range 2-8, scoring range 0-9),

If one examines the four inferences that build a case for validity in Kane's framework, most studies reported well on Scoring (mean score 2), however, all had weak rationales and discussion of Generalization (mean score 1), Extrapolation (mean score 1) and Implications (mean score 1) (**Table XVIII**). These weaknesses led to an average score for Kane's framework using the current grading scale of 6±2 (mean±SD) with a median of 5 ((interquartile 25th and 75th- 4 and 8; range 2-10, scoring range 0-12).

Table XVIII   KANE'S FRAMEWORK

Evidence Source							
	Score	Median	Min	Max	Average	SD	ICC
Scoring	0-3	2	0	3	2	1	0.80
Generalization	0-3	1	0	3	1	1	0.84
Extrapolation	0-3	1	0	3	1	1	0.71
Implications	0-3	1	0	2	1	1	0.70
Overall Score	12	5	2	10	6	2	

Evidence Source							
	Score	Median	Min	Max	Average	SD	ICC
Statement of the proposed interpretation	0-3	2	1	3	2	1	0.85
Linkage of observed performance to an interpretation	0-3	2	1	3	2	1	0.79
Linkage of the interpretation to a decision	0-3	1	0	2	1	0	0.76
Overall Score	9	5	2	8	5	1	

Outcomes: Using Cook's modification of educational outcomes for simulation that categorizes outcomes based on the domains of knowledge. skills, behavior and patient outcomes, we examined the outcomes in the 66 studies chosen. Fifteen percent of the studies tested knowledge through a combination of pre- and post-simulation testing, 48% and 15% tested time skills and behavior respectively. 66% and 70% tested process skills and behavior respectively while only 4% focused on patient outcomes. Patient outcomes were examined in three studies. Time as a measure of performance improvement dominated all studies (**Table XIX**). Total procedure time was studied in 43 reports with 80% of these studies showing a positive result. Fluoroscopy time and contrast volume, both process skills and behaviors,q were examined in 49 and 30 studies respectively with 74% and 53% of these studies reporting a positive result (**Table XX**). When various rating scales were examined, a generic scale was used in 23 studies, a procedure specific scale was used in 10 and a global rating scale was used in 7. Of the studies that used a rating scale, those employing a generic scale saw a positive response in 74%; those employing a procedure specific scale saw a positive response in 74%; and those employing a global rating scale saw a positive response in 89% (**Table XX**).

Table XIX OUTCOMES DOMAINS

Outcomes Domain	
	% of Studies
Knowledge	15%
Time skills	48%
Process skills	66%
Product skills	28%
Time behaviors	16%
Process behaviors	70%
Patient effects	4%

Table XX OUTCOMES METRICS REPORTED IN STUDIES

Outcomes Metric		
	# of Studies	Positive Study
Contrast Volume	30	53%
Total Procedure Time	43	80%
Fluoroscopy Time	39	73%
Generic Rating Scale	23	74%
Procedure Specific Scale	10	69%
Global Rating Scale	7	89%

Chapter 4 - Discussion

Summary: The literature on endovascular skills simulation is small with only sixty-six reports identified from 2000 to 2018 across multiple specialties but with most of the work concentrated in the vascular surgery field. None of the studies satisfied the design elements required in the Standards of Best Practice for Simulation with most having less than 50% of elements for a successful simulation design. Most attempted to teach complex skill set and few addressed basic skill sets or controlled for the heterogenous nature of the participants. The key simulation features were focused on identifying clinical variation, reporting the results of repetitive practice and provision of feedback. The majority of the studies used process skills and behavior (contrast volume and fluoroscopy) and under fifty percent of the studies used time skills (total procedure time) as the main educational outcome measure. Despite this, the majority of studies inferred that time to task completion was a marker of improved performance and thus, a successful simulation outcome. When one examines the overall quality of the studies on the MERSQI scale, sixty-six percent of the studies were adequate but a more detailed examination with the STARD/GRRAS tool demonstrated poor methodology with only 42% of necessary components present. A further concern is that bias was identified in the selection of the participants, which may have affected overall outcomes. No studies applied either Messick's or Kane's framework of validity and all failed to discuss key elements to demonstrate validity. Most studies referenced some validity for content as described by Messick and reported well on scoring as described by Kane.

Needs assessment for endovascular skills simulation: A major finding in this study is that none of the simulations proposed or referenced a needs assessment to justify the simulation design and the scenarios chosen. There is a paucity of needs assessments for endovascular skills across specialties to inform the prioritization and development of simulation and scenarios that should be offered to the trainees. A simple survey of attendings and fellows was performed by Woo et al (Woo, Rowe, Weaver, & Sullivan, 2012) in order to formulate a needs assessment to guide a vascular surgery skills simulation. The survey used a

Likert scale to rank 52 vascular procedures and skills. The results ranked three endovascular procedures and two open procedures as the procedures most in need for simulation. The procedures were carotid artery stenting, open repair of ruptured infrarenal aortic aneurysm, percutaneous renal artery interventions, endovascular thoracic aortic aneurysm repair and open repair of juxta-renal/supra-renal aortic aneurysm repair. In a recent Danish study, a national needs assessment was developed using a modified Delphi technique (Nayahangan et al., 2017). This study ranked basic endovascular skill sets 16th of 38 items in the needs assessment with intermediate and complex skill sets ranking even lower than the basic skill set in the final product of the modified Delphi process. Unique to this study, all the procedures were ranked according to the Copenhagen Academy for Medical Education and Simulation (CAMES) Needs Assessment formula (NAF). CAMES-NAF defines NEED as am index for the need or priority of simulation-based training for a given procedure. NEED is the *"Frequency *N * Impact * Feasibility, where Frequency is the number of procedures performed annually in Denmark, N is the number of physicians that should be able to perform the procedure, impact the impact on patients (discomfort/risk if the procedure is performed by an inexperienced doctor), and Feasibility refers to the feasibility of learning the procedure in a simulation-based environment"*. In a more recent Transatlantic consensus document (Maertens et al., 2016), twenty-four of the 26 technical skills were considered fundamental endovascular skills which corresponded to the basic and intermediate skills identified across the studies in this systemic review. However, the Transatlantic consensus demonstrated that there were significant differences were noticed between experts from Europe and the United States for five skills. Along the lines of the Danish study, a binational focused needs assessment on endovascular skills was conducted through a modified Delphi technique by Australian and New Zealand Vascular surgeons and the report laid out a list of twelve endovascular procedures that should be included in any curriculum for all vascular surgery trainees (McLachlan, Burgess, Wagner, & Freeman, 2018). Six of the 12 skills were achievable with the simulations found in this report and corresponded to basic, intermediate and complex skill sets. Importantly, carotid artery stenting, a common scenario in the majority of the studies reported, was not considered necessary. The current set of studies demonstrates that the lack of a needs assessment has directed the simulation scenarios away from scenarios that directly address the vascular surgery trainees' needs.

Quality: Fitts and Posner believe that motor skill acquisition follows three distinct stages (Fitts & Posner, 1967). The first is the "cognitive" stage where "*the learner intellectualizes the task, demonstrates understanding of the various steps and stages of the skill, such as familiarization with the various wires and catheters and learning to work with fluoroscopy"*. Once this is achieved the learner progresses to the second "integrative" stage*, which "is associated with practice, and performance is seen to flow with fewer interruptions, but the learner still needs to think in order to progress to the next procedural step"*. In the final "autonomous" stage, "*the learner has mastered the task and demonstrates fluid uninterrupted performance, while refining the finer elements of the procedure"* (Reznick & MacRae, 2006). Simulation design is key to a successful outcome. The Standards of Best Practice for Simulation allow one to determine the reported simulation design of the study (Lioce et al., 2015). Design impacts the education value to the participants, their ability to be successful and the outcomes that can be reported. The presence of a poor design directly effects validity of the simulation and a poor design will not allow an author to accurately determine the validity of their proposed simulation In the presence of poor design, quality of the methodology also suffers and adds to the difficulty in obtaining data The quality of a research paper in medical education can be assessed by MERSQI tool. Sixty-six percent of the studies achieved a passing MERSQI score of 12 or greater. While most used objective measurement the validity of the evaluation instruments was weak. This is similar to other reviews on simulation (Cook et al., 2012; See et al., 2016). Few papers have examined methodological quality in simulation. In his review on Technology-Enhanced Simulation, Cook et al (Cook et al., 2011) noted that the methodological quality, as appraised using the modified and amalgamated STARD/GRRAS tool was limited (Cook et al., 2011). In the current study using the same STARD/GRRAS criteria, we found that methodology reported in the studies was poor and where most used rating scales as a key element in assessment of the simulation, descriptions on the rating procedures, rater selection and rater training were poor across the studies. The assumption that an expert can interpret and use a scale without training and anchors was the most common flaw. The presence of flawed methodology weakens the outcomes of the studies and the possibility to prove validity for the simulations.

Bias: In a recent focused review of simulation in vascular surgery, See et al (See et al., 2016) also examined potential bias in studies, they reviewed and concluded that bias was present. See et al (See et al., 2016) did not employ a tool to determine bias, but In their opinion, biases existed either at a study or an outcome level. Selection bias was found in scenarios chosen in patient specific simulations where easier anatomy or more suitable computer tomography findings were included in the studies. It was also observed that a lack of blinding by the raters, poor or absent rate training and statistical analysis of rater interactions led to bias with regard to assessment of performance. Both the presence of and the lack of expert feedback and intervention led to confounders or even bias on assessment of improvement. Reporting biases were found with many studies focusing on reporting significant improvement in subsets of data, while the overall improvement was not significant. However, some studies reported improvement of an overall score, but the scores of subcategories were not provided. In a review of Technology-Enhanced Simulation, Cook et al (Cook et al., 2011) found methodological quality was limited. In the current study we also detected bias in participant selection with 66% of studies showing bias. Conduct of the index test, conduct of the reference test and flow did not show a high risk of bias. This coupled to issues with simulation design induces significant questions on the results and their interpretation. When the simulation design is reviewed, the lack of pre-simulation materials, the lack of pre-briefing and feedback unduly influenced the outcomes as more seasoned participants had little learning curves compared to novices. From the STARD/GRRAS methodological analysis, there were also problems with rater selection, rater training, and rater blinding which would also bias outcomes.

Validity: The majority of the studies did not reference or apply any concepts of validity. In the current report, neither Messick nor Kane were referenced in any of the papers. Cook et al (Cook et al.,

2014) performed a systematic literature search of original research that evaluated the validity of simulation-based assessment scores using two or more evidence sources. In this review only six (3%) referenced Messick's five-source validity framework, and 51 (24%) made no reference to any validity framework (Cook et al., 2014). Cook et al (Cook et al., 2014) observed that the most common concepts mentioned were: relationship to other variables (94% of studies, reported most often as variation in simulator scores across training levels), internal structure (76%, supported by reliability data or item analysis), and content (63%, reported as expert panels or modification of existing instruments). Evidence of response process and consequences were each found in <10% of studies (Cook et al., 2014). Ghaderi et al (Ghaderi et al., 2015) in a study of the validity of technical skill assessments in general surgery demonstrated that that only 3 studies of 23 assessment tools (13%) used Messick's contemporary unitary concept of validity for development of their assessment tools. Our data has shown that the unified theory was not discussed in any paper. Thus, it appears that modern framework of validity is uncommon in the simulation-based assessment within the medical education literature.

Outcomes: The current study identified several outcomes based on Cook's modification of Kirkpatrick's framework. Forty eight percent of the studies tested variables related to the time to procedure completion, and 70% tested variables related to process based on rating scales with little proven validity of the assessment tools. Repetitive practice was shown to improve time and feedback did improve process in novices, but both were not frequent tools employed across the studies. Given the paucity of modern educational theories in the simulation and curricular designs, the translational ability of the successful simulations is unknown. When compared to study designs with no intervention, Cook et al (Cook et al., 2011) reported that simulation training in a health professions educational environment is consistently associated with large effects for outcomes of knowledge, skills, and behaviors and moderate effects for patient-related outcomes. This observation is applicable to the majority of studies included in this report. In their review of vascular surgery simulation, See et al (See et al., 2016) similarly demonstrated that total procedure time and fluoroscopy time were used as surrogate performance outcomes and that repetitive practice of the scenario and/or familiarity with the simulator did significantly improve a participant's result. However, the use of contrast volume as a surrogate of process behavior did not produce as consistent a result as the other parameters, but the mechanism was unknown. An important observation was made that performing a procedure faster with less fluoroscopy and contrast use did not translate into a procedure performed for the right indication in the safest possible manner, consuming the fewest resources and with the best anticipated clinical outcome.

In a review if the vascular literature, Mitchell et al (Mitchell et al., 2014) discussed 19 articles. The checklists and global rating scales were the most commonly used metrics for objective technical skills assessment. These tools were shown to have high inter-rater reliability, construct validity, and positive user satisfaction and acceptability. In general, they were considered not very practical, as they are either procedure-specific or long (checklist of up to 62 items), making the assessment process labor-intensive, time-consuming, costly, and impractical in assessment of varying procedural skills (Mitchell et al., 2014). In a review of simulation in general and interventional radiology, Patel et al (Patel, Gallagher, Nicholson, & Cates, 2006) identified 15 articles and scored their educational outcome according to Kirkpatrick's 4 levels of achievement. Thirteen studies achieved level two of Kirkpatrick's hierarchy, with only one reaching level four of Kirkpatrick's hierarchy. The final study demonstrated no improvement in levels of achievement they also found lack of literature investigating its predictive validity and the effect on patient outcomes.

Limitations: The present study does have limitations. It is confined to the English literature and focused on the published work on endovascular simulation. It does not have access to proprietary data which commercial vendors have on their individual simulators and which may have guided their development. Many of the studies are from the same set of authors and the project uses secondary data rather than primary datasets provided by the authors.

Conclusion: The current literature on endovascular skills simulation is dominated by vascular surgery studies and is relatively small with only sixty-six reports across multiple specialties. None of the studies are considered to satisfy the design elements required in the Standards of Best Practice for Simulation. Furthermore, the research methods and reporting quality for simulation were weak and require significant refinement to allow a meaningful evaluation of outcomes. Few discussed legacy concepts of validity. When two contemporary frameworks of validity are applied, the current body of work fails to achieve sufficient rigor to be considered robust evidence to support their validity. Further work must be done to strengthen this area of assessment before widespread introduction into Graduate Medical Education curricula or professional examinations within specialties that use endovascular skills.

References

Borgersen, N. J., Naur, T. M. H., Sørensen, S. M. D., Bjerrum, F., Konge, L., Subhi, Y., & Thomsen, A. S. S. (2018). Gathering Validity Evidence for Surgical Simulation: A Systematic Review. *Ann Surg, 267*, 1063–1068.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., . . . De Vet, H. C. (2015). STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Clin Chem*, 2015.246280.

Choi, E. T., Wyble, C. W., Rubin, B. G., Sanchez, L. A., Thompson, R. W., Flye, M. W., & Sicard, G. A. (2001). Evolution of vascular fellowship training in the new era of endovascular techniques. *J Vasc Surg, 33*(2), 106-110.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ, 49*(6), 560-575. doi:doi:10.1111/medu.12678

Cook, D. A., Brydges, R., Hamstra, S. J., Zendejas, B., Szostek, J. H., Wang, A. T., . . . Hatala, R. (2012). Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. *Sim Healthcare, 7*(5), 308-320.

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med, 88*(6), 872-883.

Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., . . . Hamstra, S. J. (2011). Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *J Am Med Assoc, 306*(9), 978-988.

Cook, D. A., Levinson, A. J., Garside, S., Dupras, D. M., Erwin, P. J., & Montori, V. M. (2008). Internet-based learning in the health professions: a meta-analysis. *J Am Med Assoc, 300*(10), 1181-1196.

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract, 19*(2), 233-250.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol Bull, 52*, 281–302.

Downing, S. M. (2003). alidity: on the meaningful interpretation of assessment data. *Med Educ, 37*, 830–837.

Fitts, P. M., & Posner, M. I. (1967). *Human Performance*: Belmont: Brooks/Cole Publishing Company.

Ghaderi, I., Manji, F., Park, Y. S., Juul, D., Ott, M., Harris, I., & Farrell, T. M. (2015). Technical skills assessment toolbox: a review using the unitary framework of validity. *Ann Surg, 261*(2), 251-262.

Goodney, P. P., Beck, A. W., Nagle, J., Welch, H. G., & Zwolak, R. M. (2009). National trends in lower extremity bypass surgery, endovascular interventions, and major amputations. *J Vasc Surg, 50*(1), 54-60.

Green, S. M., Klein, A. J., Pancholy, S., Rao, S. V., Steinberg, D., Lipner, R., . . . Messenger, J. C. (2014). The current state of medical simulation in interventional cardiology: a clinical document from the Society for Cardiovascular Angiography and Intervention's (SCAI) Simulation Committee. *Cath Cardiovasc Intervent, 83*(1), 37-46.

Joint American Educational Research Association. (2014). *The Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Kane, R. L. (2006). *Understanding health care outcomes research*: Jones & Bartlett Learning.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., . . . Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Inter J Nurs Stud, 48*(6), 661-671.

Lee, J. T., Teshome, M., de Virgilio, C., Ishaque, B., Qiu, M., & Dalman, R. L. (2010). A survey of demographics, motivations, and backgrounds among applicants to the integrated 0+ 5 vascular surgery residency. *J Vasc Surg, 51*(2), 496-503.

Lewis, F. R., & Klingensmith, M. E. (2012). Issues in general surgery residency training. *Ann Surg, 256*, 553–559.

Lioce, L., Meakim, C. H., Fey, M. K., Chmil, J. V., Mariani, B., & Alinier, G. (2015). Standards of best practice: Simulation standard IX: Simulation design. *Clin Simulation in Nursing*.

Maertens, H., Aggarwal, R., Macdonald, S., Vermassen, F., Van Herzeele, I., Brodmann, M., . . . Goverde, P. (2016). Transatlantic multispecialty consensus on fundamental endovascular skills: results of a Delphi consensus study. *Eur J Vasc Endovasc Surg, 51*(1), 141-149.

McLachlan, R. H. P., Burgess, A , Wagner, T., & Freeman, A. J. (2018). A Binational Need Assessment to define thelevel of endovascular expertise required by vascular surgical trainees. *J Surg Educ, In press*.

Messick, S. V. (1995). alidity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol., 50*, 741–749.

Mills, J. L. (2008). Vascular surgery training in the United States: a half-century of evolution. *J Vasc Surg, 48*(6), 90S-97S.

Mitchell, E. L., Arora, S., Moneta, G. L., Kret, M. R., Dargon, P. T., Landry, G. J., . . . Sevdalis, N. (2014). A systematic review of assessment of skill acquisition and operative competency in vascular surgical training. *J Vasc Surg, 59*(5), 1440-1455.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann InterMed, 151*(4), 264-269.

Nayahangan, L., Konge, L., Schroeder, T., Paltved, C., Lindorff-Larsen, K., Nielsen, B., & Eiberg, J. (2017). A national needs assessment to identify technical procedures in vascular surgery for simulation based training. *Euro J Vasc Endovasc Surg, 53*(4), 591-599.

Patel, A. D., Gallagher, A. G., Nicholson, W. J., & Cates, C. U. (2006). Learning curves and reliability measures for virtual reality simulation in the performance assessment of carotid angiography. *J Am Coll Cardiol, 47*(9), 1796-1802.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *J Am Med Assoc, 298*(9), 1002-1009.

Reznick, R. K., & MacRae, H. (2006). eaching surgical skills–changes in the wind. *N Engl J Med, 355*, 2664–2669.

Schanzer, A., Nahmias, J., Korenda, K., Eslami, M., Arous, E., & Messina, L. (2009). An increasing demand for integrated vascular residency training far outweighs the limited supply of positions. *J Vasc Surg, 50*(6), 1513-1518.

See, K., Chui, K., Chan, W., Wong, K., & Chan, Y. (2016). Evidence for Endovascular Simulation Training: A Systematic Review. *Euro J Vasc Endovasc Surg, 51*(3), 441-451.

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation based assessment of clinical competence. *Adv Health Sci Educ Theory Pract, 23*(2), 323-338.

Towne, L., Wise, L. L., Winters, T. M., & Committee on Research in Education - National Research Council. (2004). *Advancing Scientific Research in Education*. Retrieved from Washington,DC:

Woo, K., Rowe, V. L., Weaver, F. A., & Sullivan, M. E. (2012). The results of a needs assessment to guide a vascular surgery skills simulation curriculum. *Ann Vasc Surg, 26*(2), 198-204.

Appendix A - Supplemental Data

Table AI BREAK DOWN OF MERSQI SCORES

Category	Variable	% present
Study design	Single group cross-sectional or single group post-test only	45%
	Single group pre-test and post-test	36%
	Non-randomized, 2 group	9%
	Randomized controlled trial	10%
Sampling	**No. of institutions studied**	
	1	90%
	2	1%
	>2	9%
	Response rate, %	
	Not applicable	0%
	50 or not reported	6%
	50-74	0%
	≥75	94%
Type of data	Assessment by study participant	12%
	Objective measurement	96%
Validity of evaluation instruments	**Internal structure**	
	Not applicable	0%
	Not reported	78%
	Reported	22%
	Content	
	Not applicable	0%
	Not reported	46%
	Reported	54%
	Relationships to other variables	
	Not applicable	0%
	Not reported	67%
	Reported	30%
Data analysis	**Appropriateness of analysis**	

	Data analysis inappropriate for study design or type of data	13%
	Data analysis appropriate for study design and type of data	85%
	Complexity of analysis	
	Descriptive analysis only	79%
	Beyond descriptive analysis	21%
Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts	9%
	Knowledge, skills	87%
	Behaviors	0%
	Patient/health care outcome	4%
Originality	Original Study	94%
	Extension of prior study	4%

Table AII STANDARDS FOR REPORTING DIAGNOSTIC ACCURACY (STARD) AND THE GUIDELINES FOR REPORTING RELIABILITY AND AGREEMENT STUDIES (GRRAS).

Study Component	Reporting element (operational definition)	% present
Title/abstract		
	Title or abstract identifying the study as an evaluation of the validity, reliability, or diagnostic accuracy of an assessment tool diagnostic accuracy of an assessment tool	12%
	Title or abstract identifying the study as focused on assessment, but not as a study of validity, reliability, or diagnostic accuracy	88%
Introduction		
	Explicit question, purpose, or hypothesis	85%
	Proposed validity argument (strategy for interpreting validity evidence to be presented)	13%
	Description of index test task	97%
	Critical review of evidence relevant to assessment of that construct	30%
Methods		
	Trainee population (eligibility criteria)	46%
	Setting (educational [e.g., simulation laboratory versus clinical)	88%
	Identification of eligible trainees (any method defined)	64%
	Sampling strategy (any method defined)	16%
	Prospective or retrospective data collection	99%
	Rationale for relationship between index and reference test	93%
	Methods/procedures for index test	100%
	Passing standard	55%
	Rater population (eligibility cri	10%
	Rater training (done or not done	10%
	Rater total number	33%

	Rater specialty	52%
	Raters blinded to trainee (done or not done)	40%
	Raters blinded to other raters (done or not done)	24%
	Raters blinded to results of reference test (done or not done)	16%
	All statistical methods defined: comparisons among groups or correlation	87%
	All statistical methods defined: reliability	3%
	Sample size calculations	1%
Results		
	Study dates	27%
	Trainee number enrolled	94%
	Trainee training level	93%
	Trainee number eligible	27%
	Flow diagram	10%
	Time interval between index and reference test	79%
	Trainee baseline proficiency: objective measurements	37%
	Trainee baseline proficiency: prior experience with that task	34%
	Central tendency (mean, median) and variability (standard deviation, range) for scores	79%
	Central tendency (mean, median) without variability	21%
	Figure (scatter plot) or table (contingency table)	10%
	Consequences of testing, adverse or beneficial	13%
	Estimate of accuracy (correlation coefficient or other)	16%
	Receiver operating characteristic (ROC) curve, sensitivity, or specificity of test	0%
	Confidence intervals for accuracy estimates	0%
	Scoring process described	88%

	Indeterminate and outlier results considered in scoring	70%
	Subgroup analyses interpreted as relating to score validity	21%
	Reliability (any)	1%
	Confidence intervals for reliability estimates	1%
Discussion		
	Clinical or Practical relevance	82%

Table AIII THE QUALITY ASSESSMENT OF DIAGNOSTIC ACCURACY STUDIES (QUADAS-2)

Participant Selection	**% Present**	
Describe methods of participant selection	73%	
Describe included participant (previous testing, presentation, intended use of index test, and setting)	60%	
Was a consecutive or random sample of subjects enrolled?	37%	
Was a case–control design avoided?	4%	
Did the study avoid inappropriate exclusions?	51%	
Could the selection of participants have introduced bias?	Low risk of bias	34%
Are there concerns that the included participants do not match the review question?	Low concern	51%
Index Test		
Describe the index test and how it was conducted and interpreted	100%	
Were the index test results interpreted without knowledge of the results of the reference standard?	69%	
If a threshold was used, was it prespecified?	46%	
Could the conduct or interpretation of the index test have introduced bias?	Low risk of bias	63%
Are there concerns that the index test, its conduct, or its interpretation differ from the review question?	Low concern	82%
Reference Standard		
Describe the reference standard and how it was conducted and interpreted	70%	

Is the reference standard likely to correctly classify the target condition?	69%	
Were the reference standard results interpreted without knowledge of the results of the index test?	48%	
Could the reference standard, its conduct, or its interpretation have introduced bias?	Low risk of bias	49%
Are there concerns that the target condition as defined by the reference standard does not match the review question?	Low concern	82%
Flow and Timing		
Describe any participants who did not receive the index tests or reference standard or who were excluded from the 2x2 table	28%	
Describe the interval and any interventions between index tests and the reference interventions between index tests and the reference standard	63%	
Was there an appropriate interval between index tests and reference standard?	67%	
Did all participants receive a reference standard?	100%	
Did all participants receive the same reference standard?	69%	
Were all participants included in the analysis?	76%	
Could the participant flow have introduced bias?	Low concern	60%

Vita

NAME:	Mark G Davies
EDUCATION:	B.A., University of Dublin, Dublin, Ireland, 1986
	M.B., B.Ch., B.A.O., University of Dublin, Dublin, Ireland, 1986
	Ph.D., University of Dublin, Dublin, Ireland, 1996
	M.D., University of Dublin, Dublin, Ireland, 1999
	M.B.A., University of Rochester, Rochester, NY, USA, 2008
CLINICAL TRAINING:	University of Dublin, Trinity College, Medicine, 1980-1988
	Royal College of Surgeons in Ireland, General Surgery, 1988-1991
	Duke University, General Surgery, 1991-1997
	University of Washington, Vascular Surgery, 1997-1999
CLINICAL EXPERIENCE:	University of Rochester, Rochester, NY, Attending Surgeon, 1999-2007
	Houston Methodist, Houston, TX, Attending Surgeon, 2008-2014
	University of Texas, San Antonio, TX, Attending Surgeon, 2015-date
PROFESSIONAL MEMBERSHIP:	American Surgical Association
	Society of University Surgeons
	Society of Vascular Surgery
	Association of surgical Education
ABSTRACTS:	Google Scholar Profile
	https://scholar.google.com/citations?user=xVe1-nsAAAAJ&hl=en
PUBLICATIONS:	Google Scholar Profile
	https://scholar.google.com/citations?user=xVe1-nsAAAAJ&hl=en