2014-12-05

Assessment of Technical Competence for Central Venous Catheterization

Ma, Irene Wai Yan

UNIVERSITY OF CALGARY

Assessment of Technical Competence for Central Venous Catheterization

by

Irene Wai Yan Ma

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

NOVEMBER, 2014

Abstract

Purpose

Central venous catheterization (CVC) is a commonly performed procedure. Technical competence in its insertion is critical for patient safety reasons. It is unclear how competence should be diagnosed. Using a unified framework of validity that outlines five sources of validity evidence, results from three published studies are presented on the use of a formative simulation-based examination, assessing for technical competence in video-recorded performances of CVC by medical trainees.

Methods

Validity evidence based on content: *study one* evaluated all available published assessment tools on CVC performances. Items on each tool were classified into competency themes.

Response process: *study two* compared the reliability measures of assessment data of 18 video-recorded performances with direct observation, rated by two independent, trained raters. Adequacy of video recordings was reviewed qualitatively.

Internal structure: *study three* used principal component analysis to assess for the dimensions assessed by an 8-item global rating scale.

Relationship with other variables: scores rated by two checklists were correlated with global assessment, confidence and number of needle attempts.

Consequences of testing: the ability of checklists to identify procedural competence was assessed.

Results

Of the 25 published checklists identified, only six (20%) assessed each of the seven competency domains. The most frequently under-represented domains were “team working” and “communication with the patient.”

Assessments between video-recorded performances and direct observations were comparable. However, wire handling was not fully captured in 13 of the 18 videos (72%). Of these, 5 (38%) were considered to have impacted rating. Drape handling was not fully captured in 17 videos (94%) and felt to be consequential to rating in 9 (53%).

Two dimensions were identified in the global rating scale: technical ability and procedural safety, accounting for 84.1% of the overall variance. For both checklists, scores correlated positively with weighted factor scores on technical ability, but negatively with scores on safety. Both checklists demonstrated lower specificity than sensitivity in the diagnosis of competence, and high checklist scores did not preclude incompetence.

Discussion

Together, these three studies presented validity evidence from multiple sources that support the use of the simulation-based examination in identifying trainees who may benefit from further training.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Kevin McLaughlin, for his guidance and feedback both on the three projects presented in this work as well as additional projects outside of this thesis. Thank you for being an excellent role model. I would like to also thank the members of my supervisory committee, Drs. Tanya Beran, William Ghali, Sylvain Coderre, and previously on my committee, Dr. Joann McIlwrick, for their support and encouragement in my work through the years. I wish to also thank my co-authors on these three publications. In particular, these publications would not have been possible without the tireless independent assessments made by Drs. Nadia Zalunardo, Nishan Sharma, and Mary Brindle. I would like to acknowledge the role of W21C in the support of all the academic work I have undertaken at the University of Calgary. I am grateful also to all my learners, including past, current, and future ones. Your thirst for knowledge inspires me to want to be a better educator. Last but not least, I wish to thank Dr. Mary Brindle for her unwavering support, patience, encouragement, and her wry sense of humour.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
α	Cronbach's alpha
ACGME	Accreditation Council for Graduate Medical Education
AUC	Area under the curve
CDC	Centers for Disease Control and Prevention
CI	Confidence Interval
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CVC	Central venous catheterization
ERIC	Education Resource Information Center
ICC	Intraclass correlation coefficient
IQR	Interquartile range
NICE	National Institute for Health and Care Excellence
OSCE	Objective structured clinical examination
p	Level of statistical significance
ROC	Receiver Operating Characteristic

CHAPTER 1: INTRODUCTION

Preamble

Central venous catheterization (CVC) is a commonly performed bedside procedure. The safe performance of this procedure is a requirement of many certifying bodies and is an important mandate given significant patient safety concerns that may arise from an imperfect execution of this complicated procedure. However, despite the clear need to identify procedural competence, it is unclear how an educator can diagnose competence in a valid and reliable manner. The ability to identify competence/incompetence is important not only for patient safety reasons, but also has implications regarding the level of supervision required of educators in the clinical setting. Using the unified framework of validity, this thesis presents three published studies that address each of the five sources of validity evidence.

Central venous catheterization (CVC)

CVC is a bedside procedure performed for the establishment of central venous access. The performance of this procedure allows for hemodynamic monitoring, volume resuscitation, and administration of inotropic agents that cannot otherwise be administered in smaller peripheral veins.[1] First introduced in the 1950s,[2] CVCs are now commonly performed in many medically-ill patients. In the United States, an estimated 15 million CVC days each year occur in the intensive care units,[3] and an estimated 20.1 million central-line days per year occur on inpatient wards.[4]

The insertion and use of CVC devices are not without risks. Infectious complications have been reported to occur in over 25 percent of patients, while mechanical complications, such as arterial puncture, hematoma, and pneumothorax,

occur in up to 20 percent of patients.[5] In addition to their direct impact on patient morbidity and mortality,[6,7] these complications have also been shown to be independently associated with increased costs and length of stay.[7,8] In one study, catheter-related bloodstream infection resulted in an attributable cost of over $10,000 and a hospital length of stay of over seven days.[7]

All medical procedures carry inherent risks and complications. However, specific to the insertion of CVCs, some of the risks from complications have been shown to be modifiable by education.

Impact of Education on Infectious Complications

It has been previously demonstrated that bundled interventions, including education on evidence-based guidelines such as hand-washing, use of full-barrier precautions, skin cleansing with chlorhexidine, avoiding the use of femoral site, and prompt removal of unnecessary catheters, led to a significant decrease in the rates of catheter-related bloodstream infection in the intensive care unit.[9,10] Results from these seminal studies suggest a possible role of education in improving clinical outcomes. However, in these studies, in addition to education, there were co-interventions, such as daily goals sheets to improve communication, creation of a central-line cart, and use of checklists to ensure adherence to infection control practices.[9,10] Therefore, the implementation of multiple co-interventions made it difficult to isolate the effect of education on patient outcomes. In a subsequent observational study, simulation-based training alone, emphasizing elements of sterility in the above evidence-based guidelines, was associated with a significantly lower catheter-related bloodstream infection rate.[11] In another observational study, education, without co-interventions, was associated with an

improvement in a behavioural outcome (e.g. an increase in the use of full-sized drapes) as well as a decrease in catheter-related bloodstream infection rates.[12] Overall, available evidence appears to support the role of education in decreasing CVC-related infectious complication rates.[13,14]

Impact of Education on Mechanical Complications

Central venous catheterization is a complex procedure with multiple steps.[15] Inexperience of the person performing CVC is a known risk factor for complications.[16-18] Indeed, it has previously been shown that physicians who have performed 50 or more CVCs have half the complication rates as those who are less experienced.[17] With appropriate education and practice, technical competence is attainable.[19-21] For mechanical complications, for example, we have previously demonstrated in a systematic review that simulation-based education, compared to control, was associated with a significant decrease in the number of needle passes and risks of pneumothorax.[22] These results argue for the role of education in decreasing mechanical complication risks.

Role of Mastery-based Learning in CVC

Traditionally, medical education has been delivered in a time-based model, also fittingly coined by some as the "tea bag model".[23] Implicit in this model is the assumption that having spent the requisite amount of time learning and "steeping" within the learning environment, the medical trainee's competence is thus attained or assumed. However, with the advent of competency-based medical education, this assumption is increasingly challenged.[24] Further, there is an increasing recognition of the critical role that assessments and frequent formative feedback play within the framework of medical education itself.[24,25] That is, in an educational program, it is no longer sufficient to only

teach. Assessment processes now become an integral component of competency-based medical education.[26]

The concept of including assessments for competency as a part of the learning process itself is not new. Mastery-based learning has long been described, whereby the educator defines at the outset what is meant by "mastery" of the subject matter and plans instructional designs and assessment procedures accordingly to reach this end goal.[27]

The existing data on the effect of mastery-based learning in health professional education in improving skills and patient outcomes have largely been positive.[21,28] For example, learners who attain pre-defined mastery skills in CVC insertion, compared to traditionally-trained learners, demonstrated a reduction in catheter-related bloodstream infection rates and mechanical complication rates.[11,29,30] In another study, mastery-based training in laparoscopic inguinal hernia repair resulted in lower perioperative complication rates, compared with standard training.[31] Thus, the role of competency-based education is increasingly recognized.

Recommendations on Education

Education on the appropriate technique for the insertion of CVC is ***strongly*** recommended by the Centers for Disease Control and Prevention (CDC).[32] Guidance on training standards for CVC has recently been made available to assist educators in selecting relevant evidence-based elements for the teaching of CVC skills, and achievement of technical competence via a final audit based on a directly-observed clinical procedure was recommended.[33]

Training to competency has been advocated by a number of national and international bodies. For example, the National Institute for Health and Care Excellence

(NICE) recommended that all those who insert CVCs using ultrasound should undertake appropriate training in order to achieve competence.[34] The CDC has also advocated training to competence. Specifically, it ***strongly*** recommends that only those who demonstrate competence for the insertion of CVC should perform the procedure.[32]

That only those who demonstrate competence should perform CVC insertions on patients makes sense from a patient safety perspective and is a difficult recommendation with which to argue. However, from an implementation perspective, this recommendation is problematic for a number of practical reasons. First and foremost, what constitutes competence? Defining competence is a difficult task and there are currently no standard methods of assessment for competence.[33] Secondly, what tools should be used to measure competence? Conflicting recommendations have been proposed, with some advocating for the use of checklists[35] while others suggest the use of global rating scale.[33] The recommendation of neither of these two approaches was based on empirical data.

Overall Purpose of Thesis

In an attempt to better address existing gaps in the literature, this thesis presents three studies to help delineate the concept of what constitutes competence in CVC insertion, what tools may be used, how the tools compare with each other, and lastly, how may competence be defined based on the use of the assessment tools.

In the spirit of competency-based education aforementioned, the setting for the assessment studies presented in this thesis is one of a *formative examination*, rather than summative. With this in mind, the primary goal of assessment addressed in this work

relates to identifying learners in need of further training, rather than identifying learners deemed competent in a summative sense.

Irrespective of whether an assessment is summative or formative, however, educators should adhere to certain principles of assessments, namely that of devising an examination that, at a minimum, yields scores which allow for valid and reliable interpretations.[36]

As such, this thesis presents three studies that will address important sources of validity evidence.[36-39] Specifically, these studies will address at least one component within each of the five sources of validity evidence in the unified framework of validity: *content, response process, internal structure, relationship with other variables,* and *consequences of testing*.[36,37] This unified framework of validity will be described in further detail in Chapter 2.

Objectives of the Three Studies

Together, the three studies aim to address the following sources of validity evidence:

1) Study one (*validity evidence based on content*): this systematic review seeks to define the construct of technical competence for CVC insertion by systematically compiling available assessment tools on CVC in the literature.[40]
2) Study two (*validity evidence based on response process*): to assess the accuracy of assessment data from video-recorded performances, this study seeks to compare the ratings of video-recorded performances with those from direct observations.[41]

3) Study three (*validity evidence based on internal structure, relationship with other variables, and consequences of testing*): Internal structure evidence was evaluated by the use of principal component analysis to assess for the dimensions captured by a global rating scale assessment tool.[42] For the assessment of evidence relating to relationship with other variables, checklist scores were compared with scores from a global rating scale, and performance scores were correlated with CVC experience, self-reported confidence, and observed number of needle attempts during CVC insertion. Lastly, the ability of checklists to identify competence/incompetence was assessed and implications regarding misclassifications discussed.

The results of these studies are by no means exhaustive or comprehensive in their attempt to address each source of validity evidence. Nevertheless, these results are intended to give educators some evidence with which to proceed with competency-based education on CVC insertion. At times, the process of validation appears to never end.[36] The results of these studies should hopefully bring educators closer to the point at which accumulating evidence supporting the interpretation of scores is at last reasonably well supported and defensible.[36]

CHAPTER 2: UNIFIED THEORY OF VALIDITY

This chapter provides a brief overview of the unified theory of validity, the specific sources of validity evidence examined by the three studies in this thesis, and lastly, why simulation-based assessment was chosen for all three studies.

Framework of Thesis: A Unified Theory of Validity

The Standards for Educational and Psychological Testing defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."[36] The current concept of validity is that of a unified approach which includes all available sources of evidence for validity.[43] Its intent is no longer to establish the accuracy of a particular score's estimate, but rather, "involves an evaluation of the overall plausibility of a proposed interpretation or use of test scores. It is the interpretation…that is validated, not the test or the test score."[43] This distinct shift, from thinking that test scores require validation to the current view that it is the *interpretation* of scores that requires validation, is the fundamental concept in the current unified theory of validity.

To support the interpretation of test scores, an evaluation of available evidence is required.[39,44,45] Under this theory, rather than evaluating different *types* of validity evidence, as was the case with the previous so-called Trinitarian approach to validation,[46] different *sources* or *lines* of validity evidence should be evaluated.[37-39,47] Specifically, sources of validity evidence now include the following: ***content, response process, internal structure, relationship with other variables, and consequences***.[38] Therefore, as much evidence as needed, from as many sources of evidence as necessary, should be

evaluated by educators who are interested in being able to identify trainees who are in need of additional training for CVC insertion.

Within this validity framework, the three studies in this thesis seek to present validity evidence, within the context of a formative simulation-based CVC technical competency assessment in an Internal Medicine residency training program.

Evidence based on Content

Content validity refers to the degree to which the items assessed are both *relevant* and *representative* of the underlying task or construct.[37,39,48,49] A construct refers to an attribute of interest that a test or examination is designed to measure.[36] In order to measure the construct of interest appropriately, test items need to be relevant and representative of the task in question. In other words, to be relevant and representative, items on a test need to cover important aspects of the underlying construct, and additionally, needs to be an adequate sample of the skill set that the test is intended to assess.[50]

Commonly used methods for establishing content validity evidence include the use of expert panels, use of existing instruments or assessment tools, use of published literature such as clinical guidelines,[47] and task or job analyses.[37] Thus, in the case of assessing for CVC technical competence, to ensure that items used in the assessment are relevant and representative, educators typically turn to expert panels or clinical guidelines for help. Not uncommonly, existing tools are borrowed from the literature for assessment purposes.

To address the overall content validity evidence of assessing technical competence for CVC insertion, ***study one*** is a systematic review that seeks to evaluate all

published assessment tools on CVC insertion in the literature.[40] In so doing, this study will comprehensively outline all previously assessed steps and domains of the procedure. In addition to the results presented in the published paper, this thesis will also present the evidence of content validity for each individual published tool.

Evidence based on Response Process

Response process refers to the evidence that supports the notion that the data are accurate and that the "sources of error associated with test administration are controlled or eliminated to the maximum extent possible."[38] In other words, the integrity of the data is evaluated.

Common examples of response process evidence includes controlling for issues relating to test security, analyzing issues with data capture, evaluating the impact of rater training, and using think aloud protocols where raters' or learners' thought processes are explored.[47]

While the most commonly sought for response process evidence is typically derived from analyses of participant or rater responses,[36,51] evidence based on response process is situated within a broader framework of data integrity and accuracy.[38] To that end, two studies in this thesis will evaluate evidence based on response processes.

First, ***study two*** seeks to evaluate the evidence of data accuracy from the ratings of performances of CVC insertion by video-review.[41] To do so, this study evaluates the method by which CVC skill assessment data was captured on video and potential methodological concerns. Finally, it compares results of ratings by video with those by direct observation.

Second, ***study three*** presents evidence based on a response process by a re-analysis of ratings of CVC performances. All performances rated as "incompetent" by raters using a global rating scale, but scored highly on the checklists, were re-analyzed for data integrity, to ensure that the scores on the global rating scale truly reflected what they were supposed to measure.[51]

Remaining Sources of Evidence

Using two published checklists on CVC insertion, ***study three*** seeks to compare the results of ratings by checklists with those by a global rating scale.[42]

Evidence Based on Internal Structure

Internal structure relates to the psychometric properties of the assessment,[38] and typically includes assessment for dimensionality via factor analysis or for item homogeneity.[36,47,52]

Study three seeks to provide evidence for internal structure by assessing for the dimensionality assessed by the global rating scale. Internal reliability of the assessment tools will also be evaluated.

Evidence Based on Relationship with Other Variables

In the absence of a gold standard tool that can measure competence with perfect accuracy and precision, correlational validity evidence is sought.[38] That is, if a given test is intended to measure competence, it *should* correlate with other measures of competence.

Commonly presented evidence on relationship with other variables include correlations with learner characteristics (e.g., training status, training level, self-reported competence), associations with performance using another measure, or in a different

setting either at the same time (i.e. concurrent validity), or at delayed date (i.e. predictive validity).[47]

For evidence on relationship with other variables, ***study three*** compares the results of rating using three measures (i.e. two checklists and a global rating scale).[42] Furthermore, this study will correlate performance measures with baseline CVC experience, self-reported confidence in the procedure, and lastly, observed number of attempts during the performance of CVC insertion.

Evidence Based on Consequences

Evidence based on consequences refers to the impact of the assessment on the examinees.[38] These consequences may include pass/fail decisions, and extends to any positive or negative consequences associated with the administration of the test, whether or not these consequences were intended or not.[38] The potential inclusion of social consequences of test use into consequential validity evidence is perhaps the most controversial aspect of this line of validity evidence,[53] but is not commonly presented in studies on simulation-based assessments.[47]

In ***study three***, consequential validity evidence will be presented based on the Receiver-Operating Characteristic curves. The implications of pass/fail decisions will be discussed further in chapter 6.

Simulation-based CVC Assessments

All three studies focused on the use of simulation-based assessments. The rationale of using a simulation-based platform for these studies is based on two primary reasons. First, simulation-based assessments allow for assessments of the procedure on an on-demand basis. Based on our previous work, attending physicians are present for only

about half of the CVCs placed,[54] suggesting that arranging for direct observation of CVC performances on patients by faculty may be problematic. Second, simulation-based assessments allow educators to study the assessment process in detail. Our learners were able to go through the procedure unchallenged, without faculty needing to intervene due to patient safety concerns.

Despite the benefits of using a simulation-based platform for assessments, there remains the concern that assessment of competence on simulators will not equate with competence on patients. None of the three studies will address this concern. Inferences regarding skill translation will instead need to be made based on others' work that directly correlated performances on simulators with performances on patients,[55-57] and correlation of performances with clinical outcomes.[22,58]

While we advocate the use of simulation-based assessments for procedural skills, the importance of workplace-based assessments on patients cannot be understated. However, the work presented in this thesis is not intended to generalize to workplace-based assessments at this time.

CHAPTER 3: STUDY ONE: VALIDITY EVIDENCE BASED ON CONTENT

Adapted from Measuring Competence in Central Venous Catheterization: A Systematic-review[40]

Background

CVC insertion is a complicated procedure with multiple steps. In the assessment of technical competence in trainees, educators may choose to adopt pre-existing tools that have previously demonstrated validity evidence to support the interpretations of scores,[37-39] or alternatively, develop new tools.[59] Where no tool exists, the need to develop one is self-evident.[59] However, even if an existing tool is available, educators may still wish to modify or develop another tool if there exists a need for a simpler or cheaper tool, or a tool that measures more appropriately the intended construct.[59]

Irrespective of the approach taken, a comprehensive understanding of the task at hand and appropriately defining the intended construct is a critical first step in the process of test design and development.[36] The degree to which the content of the examination is congruent with the intended construct will contribute to validity evidence based on content.[49]

In a previous systematic review of existing assessment tools for procedural skills in general, McKinley et al. identified seven major themes assessed by items in published assessment tools.[60] These themes included: 1) procedural competence, 2) preparation, 3) safety, 4) communication and working with the patient, 5) infection control, 6) post-procedural care, and 7) team-working. In that review, items on "infection control" and "safety" were frequently under-represented for general procedural skill assessments.[60]

Specific to the assessment of CVC skills, Evans and Dodge evaluated nine assessment tools on CVCs, published in 2009, and summarized their intended uses.[61] However, the authors did not conduct a thorough evaluation of the domains assessed by those tools. Further, there had been multiple publications both before and after 2009. Therefore, the degree to which existing assessment tools capture the intended underlying construct for CVC competence remains unknown.

Using McKinley's framework of the seven major themes on procedural competency,[60] this chapter explores to what extent these domains are being assessed by existing tools on CVC insertion. The results of this chapter will help provide some evidence of validity based on content by describing the comprehensive range of items underlying the construct of CVC competency.

Objective

The objective of this study is to systematically review existing published assessment tools for rating CVC insertion and to determine the individual steps and key competencies evaluated by these tools. Further, the content validity evidence of each existing tool will be presented.

Methods

Data sources and search strategy

We used a previously published literature search strategy that was developed with the assistance of a research librarian for our previously published systematic review of CVC education.[22] We conducted searches for relevant articles published between January 1950 and May 2010 using the following databases: PubMed, MEDLINE, Education Resource Information Center (ERIC), the Cumulative Index to Nursing and Allied Health

Literature (CINAHL), Excerpta Medica, and Cochrane Central Register of Controlled Trials. The following keywords were searched as subject headings, medical subject headings, and text words: *catheterization, central venous; catheterization; catheter$; jugular veins; subclavian vein; femoral veins*. These terms were combined with terms on education using the Boolean operator "and." Terms on education included: *education; learning, teaching;* and *teac*h$. We did not place a language restriction on the search. We performed additional hand searches for references in our included articles and relevant review articles. The initial screening of the search results for eligibility was done independently by two investigators (IM, MB), using titles and abstracts. Full-length articles were retrieved for abstracts that were felt by at least one investigator to potentially meet inclusion criteria and for abstracts that did not provide sufficient information to determine eligibility.

Selection of articles

We included primary research articles that described the assessments of CVC skills under direct observation or via video-recorded performances. Performances without any rater directly observing were excluded. For example, we did not include articles that only assessed procedural outcomes (e.g., successes/failures or complications). We excluded studies on peripherally-placed venous access devices as well as studies without an educational intervention. We also excluded articles that did not provide a description of the assessment tool in detail sufficient to replicate the tool. For all studies where only descriptions of the assessment items were reported, without provision of the assessment tool, we contacted the authors to obtain the full tool. Selection of articles at this stage was

done independently by two investigators (IM, NS). We resolved by consensus disagreements on inclusion or exclusion of articles at each stage.

Data abstraction

Independent data abstraction on baseline characteristics of each study was performed by two investigators (IM, NS) using a standardized data form (Appendix A). Information on the learner population, observers, and tools was obtained from each publication. We also abstracted information on whether or not the tool was used on patients (henceforth referred to as clinical tools) or on simulators (referred to as simulation tools).

Checklists and global rating scales

An item was considered to be part of a "checklist" if the item is an observable action. This action may be scored in a binary fashion (e.g., yes/no) or in a non-binary fashion. For example, "need for help from senior resident"[55] is an observable action and thus is considered to be a checklist item, irrespective of whether or not the authors specified the tool as a "checklist."

We defined global rating scale items as those that used a Likert scale for rating either an overall impression of the performance or on individual qualities within the performance.[62]

Classification of items into seven competency themes

Each checklist item was independently classified by two investigators (IM, MB) into one or more of the seven competency themes previously identified:[60] 1) preparation, 2) infection control, 3) communication and working with the patient, 4) team working, 5) safety, 6) procedural competence, and 7) post-procedure.

We defined "preparation" as any steps prior to the breach in patient skin (i.e. administration of anesthetics or insertion of the needle). Steps after the breach in the patient's skin but before securing of the catheter were considered part of "procedural competence." Lastly, we defined any steps including or after securing the catheter as "post-procedure," such as placement of dressings, obtaining chest radiographs, documentation of the procedure, and equipment clean-up.

Items may be classified into more than one theme. For example, an item on obtaining informed consent was classified into both "preparation" as it involves assessing for indications and contraindications for the procedure[60] as well as "communication and working with the patient," as it involves sharing information about the procedure with the patient.[60] Disagreements on classification of items were resolved by consensus.

Immediate complications are included as assessment items only if they are part of the directly-observed evaluation. For example, carotid puncture, pneumothorax, hemothorax, malignant arrhythmia, and number of needle passes are considered immediate complications. Long-term complications, such as catheter-related infections are excluded, as these "distal" outcomes may or may not be directly related to the learner's performance.

Statistical Analysis

Data were analyzed using standard parametric and non-parametric techniques. Comparisons of continuous variables between groups were performed using Student's *t*-tests. Inter-rater agreement in study selection is estimated using the kappa statistic. All analyses were performed using SAS version 9.2 (SAS Institute Inc., Cary, NC, USA) and

Stata 11.0 (StataCorp LP, College Station, TX, USA) was used for the estimation of the 95% confidence intervals (CI) for the kappa statistics.[63]

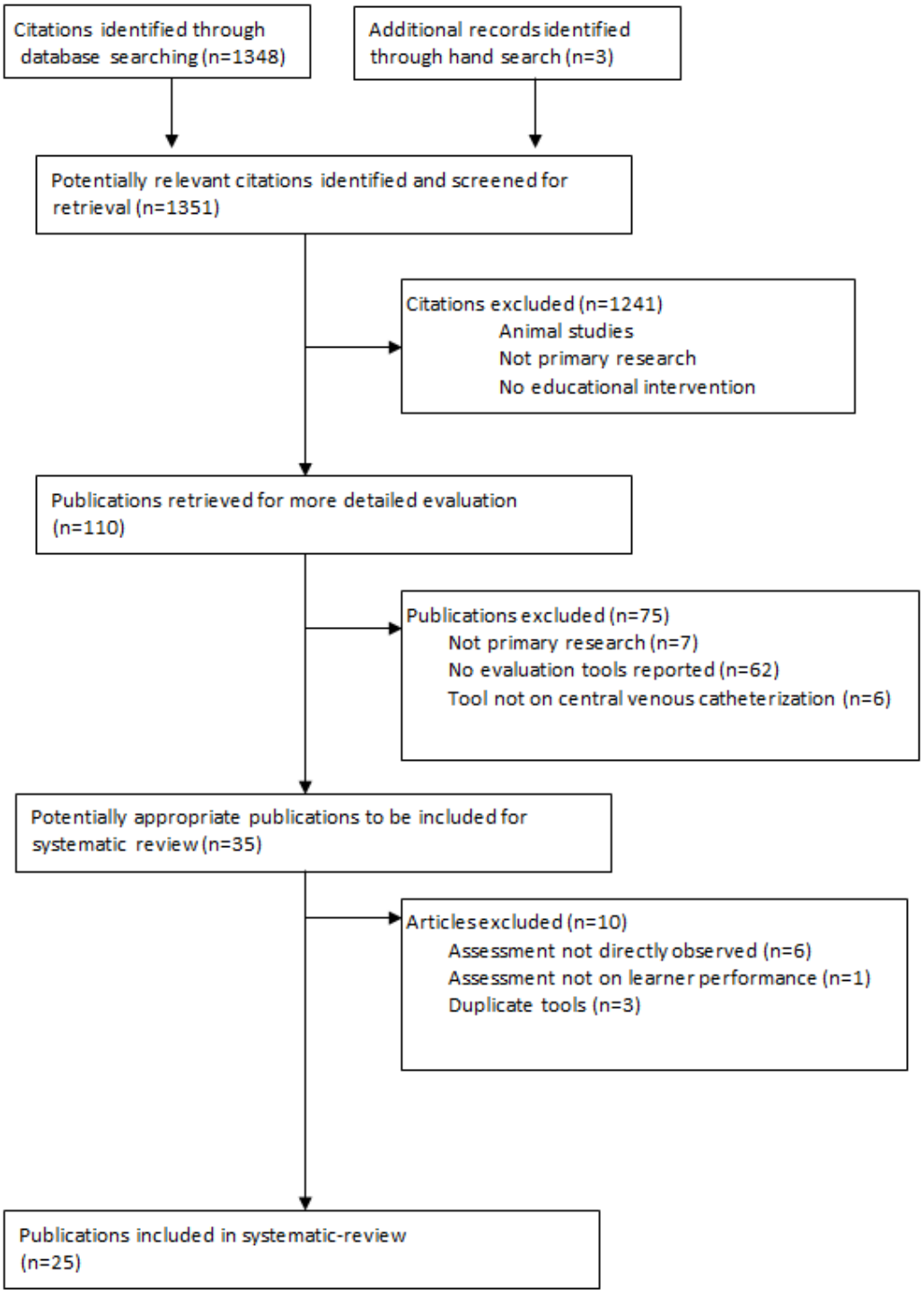
Results

Search results and article overview

Our previous search strategy from our systematic review yielded 110 articles, after excluding 1,241 articles from the initial search of 1,351 citations[22] (Figure 1). Agreement between investigators at this initial stage was high (kappa 0.87; 95% CI: 0.82 to 0.92).

From these 110 publications, 75 articles were excluded for reasons as outlined in Figure 1. Agreement at this stage was also high (kappa 0.82; 95% CI: 0.71 to 0.93). Thus, 35 articles were considered for review. Of these 35 articles, an additional ten articles were excluded (kappa 0.85; 95% CI 0.66 to 1.00). A final pool of 25 publications was included in this systematic review. Figure 1 illustrates the results of the entire study selection process.

Figure 1. Flow diagram of the study selection process



Baseline description of tools

Overall, a total of 147 items were included in the assessment tools in 25 studies (Table 1). Median number of items included per study was 17 [interquartile range (IQR) 8-22; range 2-63]. All studies reported using checklists (at least one binary item on an observable action). Only six studies reported also using global rating scales.[64-69] Other baseline characteristics of the assessment tools are listed in Table 2.

Procedural checklists

Except for two studies, checklist items were scored in a binary fashion in general. One study used a Likert scale of 1-5 (1="very unsatisfactory"; 5 = "very satisfactory") to score the seven items of observed actions in the checklist,[69] while the other study used a behaviorally anchored scale of 0-5 with a descriptor for each score to rate each of the 22 checklist items: [0 = "displays complete unfamiliarity with the step, needs visual and verbal instruction in order to perform the step ('stumped'), or omits step completely"; 5 = "executes procedure step independently, smoothly, with total confidence, and without error."][70] The remaining studies scored checklist items strictly in a binary fashion.

Thematic content of checklist items

There were 11 checklists applied to the assessments of procedural performances on simulators (simulation checklists) and 14 checklists applied to the assessments of procedural performances on patients (clinical checklists) (Table 2).

Clinical checklists had a higher percentage of items representing "preparation" and "infection control" than simulation checklists (67± 26% vs. 32 ± 26%; $p = 0.003$ for "preparation" and 60 ± 41% vs. 11 ± 17%; $p = 0.002$ for "infection control.")

Table 1. Items assessed by central venous catheterization assessment tools in the 25 studies

	Description of Items*	Evans et al (2009)	Barsuk et al (2009, Arch Intern Med)	Barsuk et al (2009, Am J Kid Dis)	Berenholtz et al (2004)	Blaivas and Adhikari (2009)	Britt et al (2009)	Carvalho (2007)	Coopersmith et al (2002)	Costello et al (2008)	Dong et al (2010)	Huang et al (2009)	Kilbourne et al (2009)	Lee et al (2009)	Lobo et al (2005)	McKee et al (2008)	Millington et al (2009)	Papadimos et al (2008)	Ramakrishna et al (2005)	Stone et al (2010)	Wall et al (2005)	Yilmaz et al (2007)	Murphy et al (2008)	Rosen et al (2009)	Velmahos et al (2004)	Xiao et al (2007)
1	Patient interaction																		1		2					
2	Verification of patient id										1													0.5		
3	Obtains consent	1	1	1							1										1		0.33	0.5		
4	Warns patient cold													1												
5	Places patient on monitor	1																								
6	Positions patient (neck at 45 degree angle)	1																								
7	Trendelenberg position	1	1	1			1				1	1													1	
8	Prepares entire neck area with chlorhexidine, cleans skin	1	1	1	1							0.33		1	1	1		1	0.5					0.5		1
9	Type of prep observed and recorded									3	1				1.5						1					1
10	Scrub time > 30 sec (or 2 min for groin line)									2		0.33														
11	Dry time >30s									1		0.33														1
12	Repeat chlorhexidine													1										0.5		
13	Uses sterile towel													1							1					
14	Applies full drape (notation toward head)		1	1	1				1	1	1					1		1			1			1		1
15	Applies half drape											1														
16	Drape, size not specified	1					1								1				0.5				0.33			
17	Indication noted																				1					
18	Identifies anatomic landmarks (site selection)	1					1					1	2				1		1				1	1	1	
19	Identifies carotid by palpation																						1			

20	Turns on machine	1																									
21	Places gel on probe	1																									
22	Appropriate gain	1																									
23	Appropriate orientation (indicator to left)	1																									
24	US to identify IJ	0.5																						0.5			
25	US to identify carotid artery	0.5																						0.5			
26	Specifies Transverse vs Longitudinal	1																									
27	IJ compressibility by ultrasound (after sterilization)										1																
28	Lidocaine into syringe	1																						1			
29	Prepares cannulation needle (accessible or attach to syringe)	1										0.25															
30	Aligns bevel											0.25															
31	Dilator easily accessible	1																									
32	Flushes ports		1	1																							
33	Clamps ports		1	1								0.25															
34	Removes brown port cap	1	1									0.25											1				
35	Prepares catheter, not otherwise specified																							1			
36	Prepares guide wire, place on drape	1																									
37	Prepares scalpel and suture	1																									
38	Establishes distance of (prepare) catheter													1													
39	Knowledge of instruments											1					1						1	1			
40	Inappropriate preparation						1										1						0.33		1		
41	Inappropriate setup						1																				
42	Places ultrasound gel into cover, sleeve	1												1													
43	Inserts vascular probe into sterile sheath	1	0.5	0.5										1										1			
44	Places rubberband 1 at tip of probe	1																									
45	Places rubberband 2 at distal cord	1																									
46	Places sterile probe onto field	1																									
47	Places sterile gel onto skin	1	0.5	0.5										1													
48	Locates carotid artery and IJ with US	1	1	1																				1			
49	General Sterile ultrasound technique										1																

50	Hands off pager																				1					
51	Washes hands				1		1		1		1.2	1			0.5	1		1			1	1		1		
52	Max sterile barrier, not otherwise specified																					1				
53	Standard precautions - face shield/mask	1	0.25	0.33	0.33				1	2	0.2	0.25			1	0.2		0.25			1			0.25		1
54	Standard precautions - hat	1	0.25		0.33					1	0.2	0.25			1	0.2		0.25						0.25		1
55	Standard precautions - gown	1	0.25	0.33	0.33				1	1	0.7	0.25			1	0.2		0.25			1			0.25		1
56	Gown tied																									1
57	Standard precautions - sterile gloves	1	0.25	0.33	1				1	1	0.7	0.25		1	1	0.2		0.25			1			0.25		1
58	Injects lidocaine	1	1	1								1							1					1		
59	Anesthetizes deeper structures		1	1																						
60	Localizes vein with anesthetizing needle																									
61	Chooses the correct needle											1														
62	Cannulates vein with finder needle (attempts)	1																								
63	Needle insertion at proper angle, trajectory						1					4	1				1						1		1	
64	Needle too cephalad												1													
65	Aspirates while advancing		0.5	0.5								1					1								1	
66	Cannulates vein with large bore needle (attempts)	1										1							1							
67	Redirects needle if unsuccessful											1														
68	Cannulates vein (specify with ultrasound)		0.5	0.5																				1		
69	Advances needle																						1			
70	Removes syringe		1	1																			1			
71	Removes syringe, specifies without moving needle											1	1													
72	Venous blood in syringe, confirm puncture	1												1									1			
73	Threads guide wire through needle (attempts)	1										1		1			1								1	
74	Advances wire (distance not specified)													1												
75	Advances wire no more than 12-15 cm		1	1																						
76	Advances wire no more than 20cm													1									1			
77	Extends skin incision with scalpel	1	1	1													0.5						1	0.5	0.5	
78	Removes needle	1												0.5			0.5						1		0.5	

79	Dilates tract with dilator	1	1	1																		1	0.5		
80	Withdraws dilator																					1			
81	CV catheter over guide wire	0.5	1	1			1							0.5			0.5						0.5	0.5	
82	Hand on wire all times	0.5	1	1										1								1			
83	Removes guide wire	1	1	1										1			0.5					1	0.5	0.5	
84	Guidewire technique, not otherwise specified																		1				1		
85	Advances catheter (14-16 cm)		1	1																		1			
86	Advances catheter (distance not specified)													1											
87	Procedural pause									1	1	1													
88	Obtains blood return (in all three ports)	1	0.5	0.5										0.5			0.5					1	0.5	0.5	
89	Flushes all ports with saline	1	0.5	0.5										0.5			0.5					0.5	0.5	0.5	
90	Heparin flush			1																					
91	Clamps each port																1					0.5		1	
92	Places clip at skin site	1																				1			
93	Sutures CVC (in four sites)	1	1	1					1		1						1					1		1	
94	Applies sterile dressing	1		1	1				1	1						1		1						1	
95	Applies dressing (sterility not specified)		1																		1				
96	Dressing type noted																				1				
97	Avoids antibiotic ointment								1																
98	Hand hygiene post-procedure														1										
99	US to confirm venous access					1					1									1					
100	Knowledge of procedure																					1			
101	Removes all sharps from site	1																					1		
102	Removes biohazards																						1		
103	CXR	1	1	1																	1	1	1		
104	Documents procedure	1																							
105	Notify catheter ok to use		1	1																					
106	Documents complications	1																							
107	Type justified																							1	
108	Type observed and recorded	--																			1				
109	Choice of site noted	--									1										1			1	1
110	Specifies choosing SC site over others								1			1													
111	Approach noted (e.g., supraclavicular/infraclavicular)	--																							
112	Side observed and recorded	--																			1				
113	Landmark technique only	--																							
114	Ultrasound localization	--																							

	landmark																									
115	Real-time ultrasound guided / bimanual use of US and needle	--																						1		
116	Landmark failure U/S guided	--																								
117	Respect for Tissue																						1			
118	Economy and smoothness of movements																1						1			
119	Instrument handling																1						1			
120	Use of assistants									1													1			
121	Flow																1						1			
122	Overall performance/ability						1					1		1			1		1							
123	Overall comfort						1																			
124	Total number of attempts insert							1					1	1			1				1				1	
125	Total number of attempts locate	1					1				1	1	1				1				1				1	
126	Successful CVC insertion by operator	1					1				1			1					1							
127	Successful CVC insertion, specify on first attempt						1																			
128	Specifies independent										1															
129	Help needed (any help)	1					1																		1	
130	Proper supervision									2											1					
131	Time to completion	1						1			1			1			1			1					1	
132	Pneumothorax	1					1														1					
133	Hemothorax	1																								
134	Hemomediastinum	1																								
135	Vessel laceration	1																								
136	Malignant dysrhythmia	1																								
137	Air embolus	1																								
138	Hematoma	1																								
139	Transient catheter malposition	1					1														1					
140	Arterial puncture with or without significant hemorrhage	1				1	1	1																		
141	Break in sterile technique/maintain sterile technique (number of times)	1	1	1	1							1						1			1					3
142	Through periosteum												1													
143	Did all other personnel follow sterile procedures?				1					1						0.2		1								
144	Number of times posterior wall penetrated					1																				

145	Responds to monitor with SVT/VT																									
146	Emergent/elective																				1					
147	Mayo stand cover																							1		

* If contents of two items contributed to one actual item on the checklist, then each item listed above is considered to have contributed to 0.5 items.

Table 2. Baseline characteristics of the 25 studies describing directly observed central venous catheterization performances

Study	Observers	Learner Population	No. of Learners observed	No. of procedures observed	Live vs Video	Sites tested (IJ = internal jugular, SC = subclavian)	Ultrasound used (yes/no)	Evaluations on patients (clinical) vs simulators	Checklist used (yes/no); no. of items	Global rating scale used (yes/no); no. of items	Additional items assessed (yes/no); no. of items
Barsuk et al (2009, J Hosp Med)	Faculty	MICU residents	28	N/A	Live and video	IJ, SC	Yes	Simulators	Yes; 27	No	No
Barsuk et al (2009, Am J Kidney Dis)	Faculty	Nephrology fellows	18	N/A	Live and video	IJ	Yes	Simulators	Yes; 27	No	No
Berenholtz et al (2004)	Nurses	ICU residents	N/A	64	Live	N/A	N/A	Clinical	Yes; 8	No	No
Blaivas and Adhikari (2009)	Faculty	Emergency medicine residents	25	25	Live	IJ	Yes	Simulators	Yes; 2	No	Yes; 1
Britt et al (2009)	Trauma fellow or critical care	Junior surgery residents	34	73	Live	N/A	N/A	Clinical	Yes; 14	Yes; 2	Yes; 1

	surgeon										
Carvalho (2007)	Medical students	Medical students	9	N/A	Live	IJ, SC	Yes	Simulators	Yes; 1	No	Yes; 2
Coopersmith et al (2002)	Nurses	Residents in surgery, anesthesiology, emergency medicine, nurse practitioner	N/A	16	Live	IJ, SC, Fem	N/A	Clinical	Yes; 9	No	No
Costello et al (2008)	N/A	N/A	N/A	N/A	Live	IJ, SC, Fem	N/A	Clinical	Yes; 18	No	No
Dong et al (2010)	Faculty and Fellow	Residents in anesthesiology, internal medicine, emergency medicine, general surgery, attending faculty	105	N/A	Video	IJ, SC	Yes	Simulators	Yes; 15	No	Yes; 3
Huang et al (2009)	Faculty	Internal Medicine residents	42	94	Video	SC	No	Simulators	Yes; 22	Yes; 1	Yes; 1
Kilbourne et al (2009)	Study authors	Surgical or emergency medicine residents	N/A	86	Video	SC	No	Clinical	Yes;6	No	Yes; 2
Lee et al (2009)	Expert reviewers	Emergency medicine residents	16	N/A	Video	IJ	Yes	Simulators	Yes; 19	Yes;1	Yes;2
Lobo et al	Infection	Medical	N/A	44	Live	IJ, SC,	N/A	Clinical	Yes;9	No	No

(2005)	control staff	residents				Fem					
McKee et al (2008)	Nurses	Pediatric anesthesiologists, surgeons, pediatric surgical staff, critical care medical staff	N/A	43	Live	N/A	N/A	Clinical	Yes; 5	No	No
Millington et al (2009)	Faculty	Medical residents	30	N/A	Video	IJ	No	Simulators	Yes; 10	Yes;5	Yes;3
Murphy et al (2008)	"Assessors"	Medical students	30	N/A	Video	IJ	No	Simulator	Yes;20	Yes;7	Yes;1
Ramakrishna et al (2005)	Cardiologists	PGY2 medical residents	20	N/A	Live	IJ	Available	Clinical	Yes;7	Yes;1	No
Rosen et al (2009)	Senior medical residents	Incoming medical residents	20	60	Live	IJ	Yes	Chickens	Yes; 22	No	No
Stone et al (2010)	Faculty	Senior medical students and PGY-1 emergency medicine residents	39	N/A	Live	N/A	Yes	Simulators	Yes; 1	No	Yes; 1
Velmahos et al (2004)	Faculty	Surgical interns	26	N/A	Live	N	No	Clinical	Yes;15	No	Yes;3
Wall et al (2005)	Nurses	"Trainees" in MICU	N/A	≥5	Live	IJ, SC, Fem	N/A	Clinical	Yes;22	No	Yes; 2
Xiao et al (2007)	Faculty	Trauma residents	50	73	Video	IJ, SC, Fem	N/A	Clinical	Yes;13	No	No
Yilmaz et al (2007)	N/A	N/A	N/A	356	Live	N/A	N/A	Clinical	Yes; 2	No	No

Simulation checklists, on the other hand, had a higher percentage of items on "procedural competence" than clinical checklists (60 ± 36% vs. 17 ± 15%; $p = 0.002$).

Representation and underrepresentation of themes

A number of checklists were comprehensive in their representation of themes (Table 3). For example, six checklists (20%) contained at least one item in each of the seven domains.[65,71-75] "Preparation" and "infection control" were assessed in most checklists: only three checklists (12%) contained no items on "preparation"[76-78] and only four checklists (16%) contained no items on "infection control."[76-79]

Other themes were less well-represented by the checklists: 13 checklists (52%) contained no items on "team working;"[66-70,76-83] 14 checklists (56%) contained no items on "communication and working with the patient;"[10,55,64,67,76-85] seven checklists (28%) contained no items on "post-procedure;"[69,76-79,82,83] seven checklists (28%) contained no items on "safety;"[10,67,69,82-85] and six checklists (24%) contained no items on "procedural competence."[80-84,86]

Table 3. Themes represented by the checklists items in the 25 studies

Study	Total No. of items	Preparation - No. items (%)	Infection Control – No. items (%)	Communication and working with the patient – No. Items (%)	Team Working- No. items (%)	Safety – No. items (%)	Procedural Competence – No. items (%)	Post-procedure – No. items (%)
Barsuk et al (2009, J Hosp Med)	27	11 (41)	6 (22)	1 (4)	1 (4)	4 (15)	13 (48)	5 (19)
Barsuk et al (2009, Am J Kidney Dis)	27	10 (37)	6 (22)	1 (4)	1 (4)	4 (15)	13 (48)	6 (22)
Berenholtz et al (2004)	8	6 (75)	8 (100)	0 (0)	1 (13)	0 (0)	1 (13)	1 (13)
Blaivas and Adhikari (2009)	3	0 (0)	0 (0)	0 (0)	0 (0)	3 (100)	3 (100)	0 (0)
Britt et al (2009)	14	6 (43)	2 (14)	0 (0)	1 (7)	7 (50)	9 (64)	1 (7)
Carvalho (2007)	1	0 (0)	0 (0)	0 (0)	0 (0)	1 (100)	1 (100)	0 (0)
Coopersmith et al (2002)	9	6 (67)	7 (78)	0 (0)	0 (0)	1 (11)	0 (0)	3 (33)
Costello et al (2008)	18	15 (83)	13 (72)	1 (6)	9 (50)	5 (28)	0 (0)	2 (11)
Dong et al (2010)	15	11 (73)	6 (40)	2 (13)	1 (7)	3 (20)	3 (20)	1 (7)
Evans et al (2005)	61	31 (51)	9 (15)	1 (2)	3 (5)	18 (30)	24 (39)	8 (13)
Huang et al (2009)	22	12 (55)	5 (23)	1 (5)	1 (5)	2 (9)	11 (50)	1 (5)
Kilbourne et al (2009)	6	1 (17)	0 (0)	0 (0)	0 (0)	2 (33)	5 (83)	0 (0)
Lee et al (2009)	19	9 (47)	6 (32)	1 (5)	0 (0)	4 (21)	9 (47)	1 (5)
Lobo et al (2005)	9	8 (89)	9 (100)	0 (0)	0 (0)	1 (11)	0 (0)	1 (11)
McKee et al (2008)	5	4 (80)	5 (100)	0 (0)	1 (20)	0 (0)	0 (0)	1 (20)

Millington et al (2009)	10	2 (20)	1 (10)	0 (0)	0 (0)	0 (0)	6 (60)	2 (20)
Murphy et al (2008)	20	4 (20)	1 (5)	1 (5)	0 (0)	6 (30)	13 (65)	3 (15)
Papadimos et al (2008)	7	6 (86)	7 (100)	0 (0)	1 (14)	0 (0)	1 (14)	2 (29)
Ramakrishna et al (2005)	7	3 (43)	1 (14)	1 (14)	0 (0)	0 (0)	4 (57)	0 (0)
Rosen et al (2009)	22	13 (59)	7 (32)	1 (5)	0 (0)	3 (14)	6 (27)	3 (14)
Stone et al (2010)	1	0 (0)	0 (0)	0 (0)	0 (0)	1 (100)	1 (100)	0 (0)
Velmahos et al (2004)	15	5 (33)	2 (13)	0 (0)	1 (7)	2 (13)	7 (47)	3 (20)
Wall et al (2005)	22	17 (77)	9 (41)	2 (9)	2 (9)	4 (18)	2 (9)	3 (14)
Xiao et al (2007)	13	13 (100)	12 (100)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Yilmaz et al (2007)	2	2 (100)	2 (92)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Global rating scales and additional items assessed

Only six studies reported the use of global ratings scales (Table 4),[64-69] all of which were used in conjunction with checklist items. The median number of items assessed by the global rating scale was 2 (IQR 1-5; range 1-7).

Additional items evaluated by the assessment tools frequently included continuous measures such as the number of attempts and time taken to perform the procedure (Table 5).

Validity and reliability evidence for the assessment tools

Inter-rater reliability was reported for 12 (48%) of the studies,[65-68,70-73,75,78,79,82] reporting a range of reliability coefficients and absolute agreement (range between 0.35 for reported weighted kappa statistic[67] to 0.97 for reported overall agreement).[73]

Only 12 studies (48%) specified the process used to support validity evidence based on content.[55,65,66,70-75,79,81,86] The most commonly cited sources for content validity were the use of literature review, reported by nine studies,[65,66,71-73,75,79,81,86] and the use of expert review/expert panel, also reported by nine studies.[55,65,66,70-73,75,86] In one study, a Delphi approach was described for obtaining expert consensus,[65] while a systematic approach to checklist construction[87] was reported by two studies.[71,72] One study incorporated information from local workflow assessments,[74] while a cognitive task analysis approach was used by another study.[55] Pilot testing of the assessment tools were described only in three studies,[10,66,74] although the degree to which the piloting contributed to content validity evidence was not specified in these studies. Two additional studies utilized previously published tools.[67,68] The remaining 13 studies did not cite any evidence for content validation.

Table 4. Items used in global rating scales

Study	No. of Global Rating Scale Items	Items	Scale Used
Britt et al (2009)	2	Resident comfort; Resident ability	1-5
Huang et al (2009)	1	Overall performance	Anchored 1-5 (1 = "unable to complete procedure without assistance", 3 = " demonstrates essential skills to complete procedure", 5="demonstrates mastery of procedure skills")
Lee et al (2009)	1	Overall performance	1-7 (Poor to excellent)
Millington et al (2009)	5	Time and motion; Instrument handling; Flow of operation and forward planning; Knowledge of instruments; Overall rating	1-5 behaviorally anchored scales[1] 1-5 (1="overall does not meet expectations"), 5 ="superior, exceeds expectations")
Murphy et al (2008)	7	Respect for tissue; Time and motion; Instrument handling; Knowledge of instruments; Use of assistants; Flow of procedure and forward planning; Knowledge of specific procedure	1-5 behaviorally anchored scales[1]
Ramakrishna et al (2005)	1	Overall perception: Resident is capable of *independently* performing central line procedures	1-5 (1= "Strongly disagree", 3 = "Neutral", 5 = "Strongly agree")

[1] Global rating scale based on scale from Reznick R, Regehr G, MacRae H et al. Am J Surg 1997; 173(3):226-30.

Table 5. Additional items assessed by the assessment tools

Study	No. of additional items assessed	Items
Blaivas and Adhikari (2009)	1	Number of times posterior wall penetrated
Britt et al (2009)	1	Average sticks to cannulation
Carvalho (2007)	2	No. of attempts required to cannulate the vessel; Time from skin penetration to successful guidewire insertion and needle removal
Dong et al (2010)	3	Number of venipuncture attempts; Number of skin entries; Procedural time (from initial greeting of the 'patient' until successful catheterization)
Evans et al (2010)	2	Total number of attempts to cannulate vein with large bore needle; Time to completion
Huang et al (2009)	1	Number of passes
Kilbourne et al (2009)	2	Number of insertion attempts ; Number of unsuccessful failure
Lee et al (2009)	2	Number of attempts; Time needle touches skin and time vessel successfully puncture
Millington et al (2009)	3	Number of attempts to locate the vein; Number of attempts to insert the catheter; Total time for procedure
Murphy et al (2008)	1	Time taken to complete the procedure
Stone et al (2010)	1	Time from first synthetic skin puncture until "flash";1
Velmahos et al (2004)	3	Number of attempts to locate the vein; Number of attempts to insert the catheter; Time to complete procedure.
Wall et al (2005)	2	List all sites where insertion was attempted; How many different needle sticks did the patient receive (number of skin breaks)?

Discussion

This systematic review identified 25 studies reporting the use of assessment tools for CVC performances. All of these tools used at least one item that was scored in a checklist fashion and only six studies reported also using a global rating scale. By more systematically and comprehensively evaluating the available literature, our results extend those of a previous report on the frequency with which checklists were used for the assessment of CVC performances.[61]

Our study identified that only 20% of the assessment tools incorporated at least one item in each of the seven key procedural competency domains and that the majority of tools did not assess for competency in the domains of "team working" and "communication and working with the patient."

In an effort to improve clinical outcomes through the use of competency-based training using simulation, educators need to be mindful of assessing all the relevant domains to the construct of interest. Therefore, domains that have potential patient safety implications, such as "safety," and "infection control," should be included.

Tools are frequently created with specific purposes in mind. Thus, our finding that tools used clinically contained more items in "preparation" and "infection control" than tools used in simulation-based observations likely reflects the intended purposes of the assessments. Simulation-based checklists containing more items on "procedural competence" than clinical checklists is likewise a reflection of the specific purposes of the simulation-based examinations, whereby the learner's technical skills can be assessed unchallenged on a simulator. For some simulation-based assessments for CVC, focused primarily on technical competence, items on "preparation" may or may not be relevant to the intended construct in question. For example, gathering equipment may not be a relevant item if the simulation-based examination has all the

relevant equipment already at the station. However, arguably, items on "infection control" should be integral to any CVC technical assessments, as each step executed in the technical procedure needs to be performed in a sterile manner; otherwise, infectious complications may ensue.[5]

Limitations

There are some limitations in this systematic review that impact on the interpretation of our study's conclusions. First, despite our systematic review including only publications that included an educational intervention, the assessment purposes of the studies were not uniform. Tools designed to be used by nurses for the sole purpose of documenting infectious risks clinically or tools designed for the purposes of assessing only technical skills on simulators are unlikely to be as comprehensive as tools designed to assess for holistic competence in procedural skills.[88] Indeed, our results suggest that clinical checklists were more focused on steps involving preparation and infection control than tools used on simulators, while simulation-based tools were more focused on procedural technical competence itself. Therefore, the contextual features of each published tool are important to recognize. Second, although we contacted multiple authors to obtain the actual assessment tools, only a number of authors complied.[74,80,86] Therefore, a number of studies were excluded because of a lack of response from the authors. Third, because our initial literature search was based on our previously published systematic review,[22] we included only studies up to and including May 2010. Fourth, while we presented items assessed by checklists and global rating scales, we have not directly compared the two assessment modalities, and therefore cannot directly advocate for the use of one over the other. This issue will be further examined in Chapter 5. Lastly, while we have published a comprehensive list of possible steps for the assessment of CVC competence, we do not advocate

for the use of all 147 items for assessment purposes. Firstly, in any given examination setting, some of the items will not be applicable. Therefore, those construct-irrelevant items should be removed. Secondly, even if all items are applicable, validity evidence for the use of this 147-item tool has not been previously reported. Therefore, any interpretations of scores resulting from the use of this tool are at present unsupported. Lastly, the use of such a lengthy and cumbersome tool is likely not feasible, and further, may run the risk of rewarding thoroughness in technique rather than actual competence.[49,89]

Despite these limitations, this study has a number of strengths. First, this study systematically evaluated all published tools on CVC insertion. By providing a comprehensive evaluation of existing tools on CVC assessments, this study, to our knowledge, is the first to compile comprehensively all potential relevant steps for CVC insertion (Table 1). However, the degree to which these individual items are relevant to the underlying intended construct of CVC performance will still need to be determined. For example, if the intended construct is simply technical competence, then the item on obtaining a chest radiograph post-procedure may be considered by some to be construct-irrelevant. If, on the other hand, the intended construct of interest is to assess for procedural competence in a comprehensive and holistic manner, then the exclusion of this item on the assessment tool may well result in construct underrepresentation.

The availability of this comprehensive list of steps will allow the educator to gauge the degree to which any given tool "fully and sufficiently represent the targeted domain."[49] A second strength of this study is that by comprehensively compiling all available tools, this study presents to the reader a range of sources of validity evidence based on content. For example, rather than relying solely on expert panels, literature review, cognitive task analysis, or workflow assessments, this study presents results based on all of the above sources. This allows the

educator to choose a desired tool based on evidence, or alternatively, re-create a new tool with the above evidence in mind.

CHAPTER 4: STUDY TWO: VALIDITY EVIDENCE BASED ON RESPONSE PROCESS

Adapted from "Notes From the Field: Direct Observation versus Rating by Videos for the Assessment of Central Venous Catheterization Skills"[41]

Background

Data for assessment purposes need to be accurate. Thus, educators need to control or minimize sources of error as much as possible.[38] Evidence that contributes to the control or minimization of errors contributes to *response process validity*, and examples of such includes analyses of rater disagreements, problems with video collection, and think aloud protocols that analyze raters' thought processes that occur during rating.[47] While evidence based on response processes are most commonly derived from analyses of participants' or judges' responses,[36] this chapter is devoted to a less commonly evaluated source of evidence. Specifically, this chapter will evaluate issues surrounding video recording[47] for the assessment of CVC technical skills as it relates to data accuracy. Additional evidence based on response process derived from judges' ratings will be evaluated further in Chapter 5.

The purported benefits of video recording

Video-recorded performances are frequently used to assess learner performances.[90,91] Unlike direct observations where the assessor must be present during the performance of the procedure, ratings by video can be done at a time and location convenient to the assessors and allow them to rewind and fast-forward segments of the performances as necessary.[91-93] Further, depending on the camera angles, video-recorded performances allow blinded assessments by not

including identifying features of the trainees,[94] thereby eliminating examiner familiarity as a source of bias in the rating of performances.[95-98]

However, video-recorded performances are not without potential disadvantages. For example, poorly recorded performances without sound or those that do not capture the relevant details of complex procedures may impair the ability to accurately rate performances. For example, in one study on laparoscopic cholecystectomy, the correlation between ratings on video-recorded performances and direct observation was found to be poor.[99] The authors noted that the lack of audio information on the video-recorded performances in that study resulted in a loss of crucial information, making operative movements difficult for the raters to interpret.[99] A second potential disadvantage to the use of video-recorded performances includes the added costs and complexity associated with equipment issues and AV storage needs. Lastly, validity evidence supporting tools designed for direct observation may not be transferrable to video-recorded performances.[100] As can be seen from our systematic review (Table 2, Chapter 3), there are more tools developed for the ratings of performances under direct observation than those for the ratings of video-recorded performances.

Use of video recordings for assessment purposes is not uncommon. However, with a few notable exceptions,[93,99,101-103] evidence for response process in terms of the adequacy of video-recordings is seldom discussed. Even in studies where the recording adequacy was discussed, it was seldom discussed in detail. For example, in a study by Aggarwal et al., the authors noted that all 53 videos of the procedure were of "adequate quality for video-based analysis."[102] Little additional information was provided. In another study by Kneebone et al., camera resolution was noted to limit the ability to observe details of procedural techniques.[101] However, the impact of this limitation on the accuracy of assessments was not further delineated in that study. In the

study by House et al., three performances were excluded from analysis due to sound issues.[103] In the study by Scott et al., in addition to the lack of audio information, additional technical problems such as blank tapes, dark tapes, or tapes that ended prematurely were noted in 16% of cases.[99] Lastly, Beckmann et al.'s study found that 54% of videotapes was unsatisfactory due to technical problems and poor camera work.[93]

The majority of the available studies on video-recorded assessments focused primarily on feasibility issues and psychometric analyses of assessments using video recordings, but no results from a comparison group were provided.[91,93,102,104-106] For studies with a control group, there was little information on the limiting aspects or lack thereof with respect to technical issues.[92,100,103,107] While correlation between direct observation and video-recorded assessments were generally found to be high or substantial,[92,103,107] this finding was not universal.[99] Further, while one study found that assessments by direct observation lacked construct validity,[106] another study found that video-recorded assessments lack construct validity.[99]

For studies on CVC, 14 studies evaluated learners by direct observation, while seven studies used video-recorded assessments.[40] Only two studies used both methods of assessments.[71,72] However direct comparisons between the two methods were not performed in those studies. Therefore, how these two methods compare in the assessment of CVC remains unknown. For video-recorded assessments, the accuracy of assessment ratings is limited by the accuracy of the available information captured on video. It may be tempting to infer that the disparate results in the literature on the correlations and construct validity between direct observation and video-recorded assessments may suggest mixed results also for CVC assessments. However, comparisons between direct-observation and video-recorded assessments to date have not been empirically examined for CVC. Therefore, this study seeks to compare

assessments on video-recorded performances with those by direct observations and in so doing, highlight some important technical issues on video-recording techniques in the assessment of CVC performances. In this chapter, more detailed analyses comparing the two methods will be presented than in the original published paper from which this chapter is derived.[41]

Methods

Two trained assessors independently assessed performances of CVC by direct observation using simulators (Laerdal IV Torso; Laerdal Medical Corp, Wappingers Falls, New York). These procedures were performed by 34 first year Internal Medicine residents who provided written consent, a cohort that will be described in further details in Chapter 5.[42] In short, participants underwent a 2-hour CVC training. Pre- and post-training, participants filled out a demographic survey (Appendix B). Post-training, participants underwent a formative simulation-based procedural examination.

From this original cohort of 34 examinations, we included only the 18 performances whereby two assessors were available for both direct observation and video-recorded performances, in order to allow for the assessment inter-rater reliability. Performances where only one assessor was available were excluded.

This study was approved by the university ethics review board (Ethics ID H08-01739).

Video-recording

Each procedural performance was video-recorded using Panasonic 3CCD PV-GS300 miniDV camcorders (Panasonic Corporation of North America, Secaucus, NJ). Video-recording was done in a blinded fashion, with no personal identifying information was recorded. Sound was recorded and the participants were asked to talk through the procedural steps. A camera was placed on a tripod, recording from the mannequin's right-hand side irrespective of participants'

handedness. The recording was performed in a static fashion, with the focus of the field upon the right internal jugular vein on the simulator, capturing also the upper torso of the mannequin, the participants' gloved hands and gowned arms, and the standardized procedural kits (Arrow® multi-lumen central venous catheterization kit with Blue FlexTip®, Reading, PA). All participants were instructed to perform a right internal jugular CVC by landmark technique as they would in real-life, without externally imposed time limits on the procedure.

Assessors and training

Assessors were faculty members with a minimum of three years of experience in simulator teaching and assessment and ten years of clinical experience in performing CVCs (IM, NZ). Each assessor was trained for three hours on the use of each of the two assessment tools used in this study (see below on *Assessment tools*). Training was done using select video-recorded performances that represented a range of performances from poor to excellent. After training, the intraclass correlation coefficients (ICC) of the overall scores for the two assessors were >0.80. The same two assessors assessed all performances by direct observation and the same two assessors also assessed all video-recorded performances. That is, all performances (direct observation and video-recorded) were rated using two tools by the same two assessors.

Assessment tools

Two assessment tools were used in this study: a global rating scale and a checklist.[42]

Global Rating Scale

The global rating scale used in this study is an eight-item scale, with an additional ninth summary item on "overall ability to perform procedure" (Appendix C). The eight items were adapted from two previously published global ratings scales: the Direct Observation of Procedural Skills from The Foundation Programme and the Objective Structured Assessment of

Technical Skills.[109] Only items in the original scales that are applicable to our simulator examination were used. This assessment tool was piloted by two assessors (IM, RH) on seven trainees. The rating scale was then modified and chosen based on group consensus. The summary item was a 6-point Likert scale with descriptive anchors ranging from 1 = "not competent to perform independently" to 6 = "above average competence to perform independently."

Pass/fail decisions are based on the summary item: a rated overall ability of "borderline competence" or lower is considered a failure, while those rated as "competent to perform independently" or higher are considered to have passed.

Checklist

The binary checklist used in this study is a 10-item checklist (Appendix D) that is adapted from a previously published checklist, constructed based on cognitive task analysis.[55] Items on the original checklist deemed not applicable to our simulator examination were removed. The overall checklist score was calculated as the number of completed items divided by the total number of items, presented as a percentage.

Reliability and additional validity evidence for these tools will be described in further detail in Chapter 5.

Minimization of bias

To minimize the extent to which one tool may have systematically influenced the rating of the subsequent tool, during both direct observation and analyses of video-recorded performances, for each assessor, 50% of the performances were rated first using the global rating scale, while the remaining 50% of the performances were rated first using the checklist.

To minimize assessor recall of video-recorded performances, two methods were chosen: First, video-recorded performances were assessed a minimum of 6 months after the directly-observed performances. Second, each assessor evaluated the entire pool of 34 video-recorded performances in random order, rather than only the 18 performances included in this study.

Analyses of video-recording adequacy

Video recordings were reviewed qualitatively by one investigator (IM) for the following: adequacy of sound quality, evidence of missing video-segments, ability to visualize sterility breaches and duration of time any necessary portions of the procedure is out of view. For videos where portions of the procedure were deemed inadequate, two investigators (IM, MB) independently reviewed those portions to determine if the missing segments may have impacted upon the assessments. Disagreements were resolved by consensus.

All discrepancies of assessments between direct-observation and video-recorded assessments were reviewed by two investigators (IM, MB). ***Leniency*** during direct observation or during video-review was inferred if no clear explanation accounted for the rating discrepancies. ***Inattention during video-review*** was inferred if the video clearly demonstrated the presence or absence of an action while the rating on the relevant video-review checklist item reflected otherwise. ***Inattention during direct observation*** was inferred if the video clearly illustrated the presence or absence of an action while the relevant direct-observation checklist item reflected otherwise.

Statistical analysis

Validity evidence on internal structure (reliability) will be presented along with evidence on response process. Inter-rater reliability was evaluated using intraclass correlation coefficient (ICC) and Cohen's Kappa.[63] Paired *t*-test was used to test for significance of differences in

paired continuous data while McNemar's test was used to test for significance of differences in paired dichotomous (pass/fail) data. Trends in differences between the two methods of assessments were evaluated using the Bland-Altman analysis, which plots the difference in scores between the two measures on the y-axis, and the mean of the results of the two measures on the x-axis.[110] Mean differences exceeding that of the mean difference ± 1.96 times the standard deviation of the differences are considered significant.

All analyses were performed using SAS version 9.1 (SAS Institute, Cary, North Carolina) and PASW Statistics software, version 18.0 for Windows (PASW, IBM Corporation, Somers NY). Estimations of 95% CI for kappa statistics were made using Stata 11.0 (StataCorp LP, College Station, TX).[63]

Results

Adequacy of video-recorded assessments

Sound quality was deemed adequate in all 18 video-recordings. Recording issues led to the failure to capture the beginning of the video in two (11%) assessments (missing preparation phase due to a delay in turning on the camera) and failure to capture the end of the video in one (6%) assessment (missing post-procedural care due to the end of the tape occurring earlier than expected).

There were difficulties capturing the entire wire handling process in 13 (72%) of the videos, with the wire being out of view for all videos for a median of 11.5 seconds (IQR 0 – 16.5 seconds). For only the 13 videos in which the wire was not captured, the wire was out of view for a median of 15.3 seconds (IQR 12.2-19.2 seconds). Out of these 13 videos, inability to visualize the entire wire was felt to be consequential for the assessment of wire handling in five

of the videos (38%) because hand positioning were easily ascertained in the remaining eight (62%) videos.

The camera field of view did not include the identity of any of the participants.

Because of the close proximity of the drape handling to the participants, handling of the drapes were frequently out of view (n=17; 94%). The drape was out of view for a median of 7.7 seconds (IQR 4.5 – 10 seconds). Out of the 17 videos in which the drape handling was out of view, 9 of these (53%) were felt to have had a potential impact on scoring. For example, impact on scoring was deemed to be of potential consequence to the assessment if the investigators felt that the portion of drape handling that was out of view may have revealed breaches of sterility that were otherwise not observed in the video. Impact on scoring was deemed unlikely to be of consequence if the investigators were able to ascertain that the portion of the drape that was out of view was unlikely to have made contact with non-sterile equipment or surfaces.

Differences between video-recorded assessment scores and direct observation scores

There were no significant differences in the summary global rating score between video-recorded assessment and direct observation (2.7 ± 0.9 *vs*. 2.8 ± 0.8, $p = 0.40$, where 3 = "competent to perform independently", Table 6). However, there were two items whereby the two modalities rated the same performance differently: appropriate preparation of instruments pre-procedure (3.0 ± 0.2 for video-recorded assessment and 3.2 ± 0.4 for direct observation, $p = 0.006$); time and motion (3.3 ± 0.5 vs. 3.1 ± 0.6, $p = 0.03$, respectively).

There were no significant differences in the summary checklist scores between the two modalities (91.6 ± 11.5 vs. 90.6 ± 11.7%, respectively, $p = 0.28$, Table 6).

Table 6. Agreement between video-recorded assessment scores and direct observation scores

	Video-recorded assessment score (mean ± SD*) or N (%)	Direct observation score (mean ± SD*) or N (%)	P-value
Global Rating Scale Items			
Appropriate preparation of instruments pre-procedure	3.0 ± 0.2	3.2 ± 0.4	0.006
Appropriate analgesia	2.6 ± 0.7	2.7 ± 0.9	0.32
Time and motion	3.3 ± 0.5	3.1 ± 0.6	0.03
Instrument handling	3.3 ± 0.6	3.1 ± 0.7	0.35
Flow of procedure and forward planning	3.2 ± 0.5	3.1 ± 0.5	0.57
Knowledge of instruments	3.1 ± 0.5	3.2 ± 0.5	0.49
Aseptic technique	2.8 ± 0.6	2.8 ± 0.6	0.80
Seeks help where appropriate	1.3 ± 1.4	1.5 ± 1.5	0.28
Overall ability to perform procedure	2.7 ± 0.9	2.8 ± 0.8	0.40
Overall pass percentage	N=12 (61%)	N=27 (75%)	0.10
Checklist Items†			
Prepared site appropriately	N=28 (82)	N=38 (82)	1.00
Identified proper landmarks	N=34 (100)	N=34 (100)	NA‡
Inserted needle at correct angle	N=31 (86)	N=30 (83)	0.56
Aspirated while inserting needle	N=36 (100)	N=36 (100)	NA‡
Inserted guidewire appropriately	N=36 (100)	N=35 (100)	NA‡
Withdrew needle and incised skin	N=32 (89)	N=32 (89)	NA‡
Inserted catheter and withdrew wire	N=33 (94)	N=32 (91)	0.32
Aspirated blood and flushed ports	N=30 (94)	N=30 (94)	NA‡
Occluded ports	N=25 (78)	N=23 (72)	0.15
Secured catheter	N=31 (91)	N=31 (91)	1.00
Overall checklist score	91.6 ± 11.5%	90.6 ± 11.7%	0.28

* SD denotes standard deviation

† Denominator not consistently 36 due to missing data

‡ Not applicable (no discordant pairs)

Comparing agreement of pass/fail decisions made using the global rating scale between video-recorded assessments (61% pass) and direct observation (75% pass, Table 6) showed moderate overall agreement:[111] kappa = 0.44 (95% CI 0.14 – 0.73).

There were no significant differences in the pass/fail decisions (McNemar's $S = 2.78$; $p = 0.10$). Bland-Altman plot analyses suggest good agreement between direct-observation and video-recorded assessments for both the summary global rating scores and summary checklist scores (Figure 2 and Figure 3).

Figure 2. Agreement between direct-observation and video-recorded assessment for summary global rating scale scores.

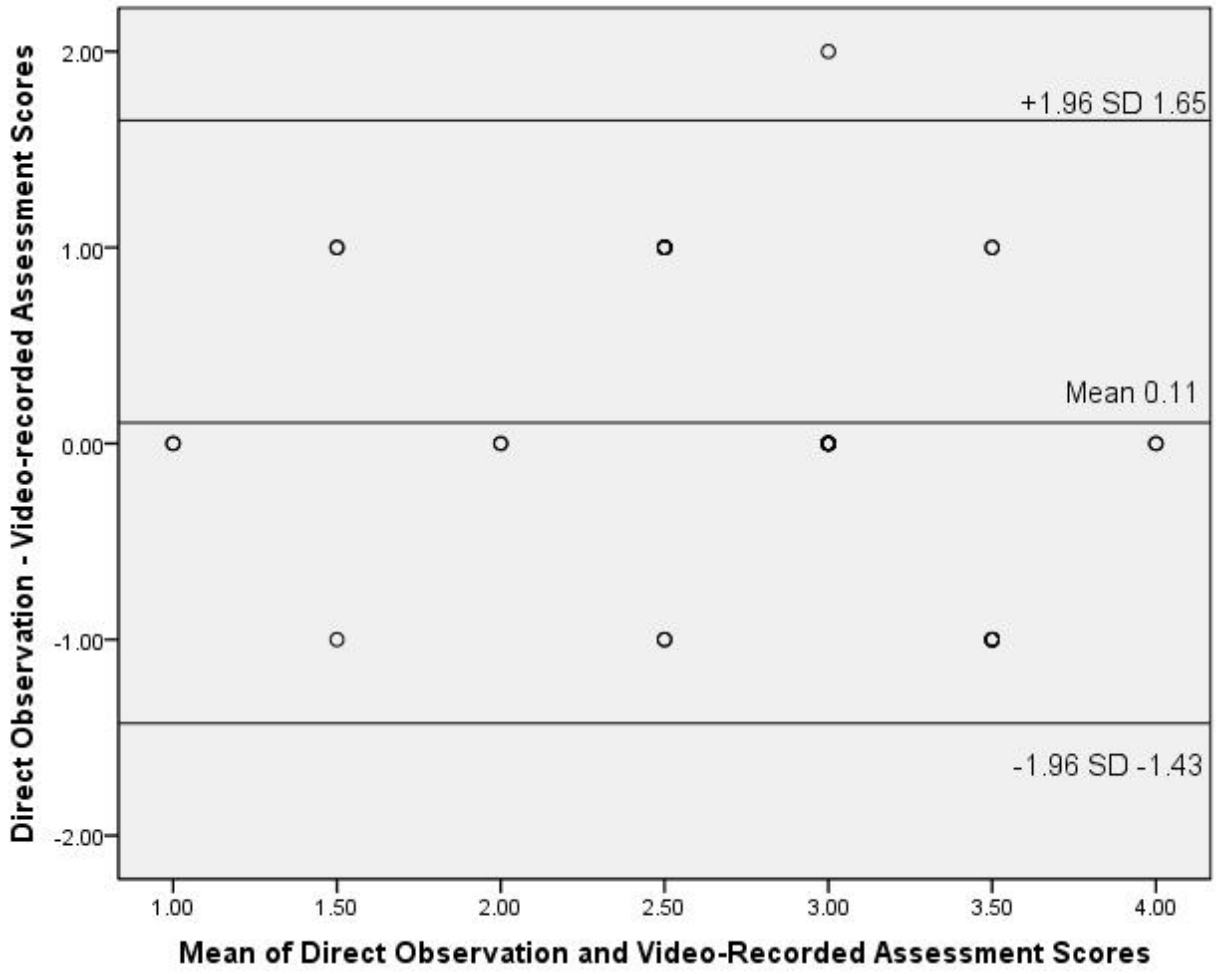
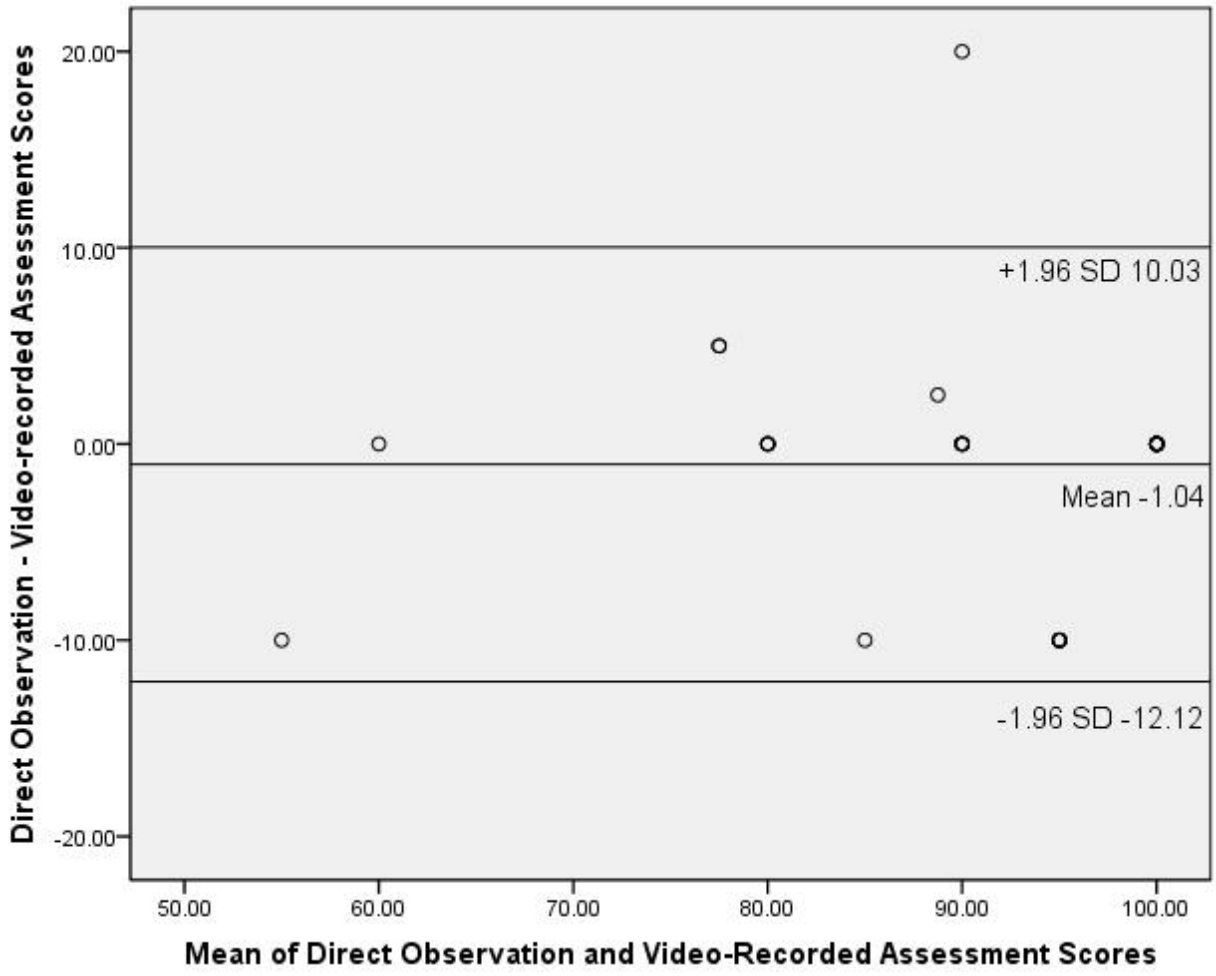


Figure 3. Agreement between direct-observation and video-recorded assessment for summary checklist scores.



Inter-rater reliability of performances rated by video-recorded assessments and direct observations

For most of the individual items on the checklist, ratings for video-recorded assessments demonstrated higher inter-rater reliability than ratings for direct observation (Table 7). However, inter-rater reliability for the overall checklist score and for the overall assessment on the global rating scale was higher for direct observation than for video-recorded performance assessments.

Table 7. Inter-rater reliability of the two assessment methods: video-recorded assessment and direct observation

	Inter-rater reliability of Video-recorded Performances (ICC)	**Inter-rater reliability of Direct Observation (ICC)**
Global Rating Scale Items		
Appropriate preparation of instruments pre-procedure	Perfect agreement	0.21 (-0.30-0.61)
Appropriate analgesia	0.69 (0.34-0.88)	0.78 (0.34-0.93)
Time and motion	0.01 (-0.47-0.47)	0.55 (0.12-0.81)
Instrument handling	0.43 (0.01-0.73)	0.47 (0.03-0.76)
Flow of procedure and forward planning	0.12 (-0.39 – 0.55)	0.49 (0.06-0.78)
Knowledge of instruments	0.22 (-0.30-0.62)	0.35 (-0.12-0.70)
Aseptic technique	0.75 (0.45-0.90)	0.32 (-0.18-0.68)
Seeks help where appropriate	0.67 (0.31-0.86)	0.67 (0.32-0.86)
Overall ability to perform procedure	0.71 (0.39-0.88)	0.87 (0.68-0.95)
Checklist Items	**Inter-rater agreement of Video-recorded performances (Kappa)**	**Inter-rater agreement of Direct Observation (Kappa)**
Prepared site appropriately	Perfect agreement	0.61 (0.09-0.89)
Identified proper landmarks	Perfect agreement	Perfect agreement
Inserted needle at correct angle	Perfect agreement	0.64 (0.07-0.93)
Aspirated while inserting needle	Perfect agreement	Perfect agreement
Inserted guidewire appropriately	Perfect agreement	NA (no variability)
Withdrew needle and incised skin	Perfect agreement	Perfect agreement
Inserted catheter and withdrew wire	Perfect agreement	0.64 (0.07-0.93)
Aspirated blood and flushed ports	Perfect agreement	Perfect agreement
Occluded ports	0.48 (-0.03-0.80)	0.85 (0.53-1.00)
Secured catheter	0.64 (0.06-0.93)	0.64 (0.07-0.93)
ICC of checklist score	0.81 (0.47-0.93)	0.88 (0.72-0.95)

Discrepancies between direct-observation and video-recorded assessments

There were a total of 26 discrepancies between direct observation and video-recorded assessments in checklist items. Upon review of these discrepancies, 16 of these discrepancies (62%) were attributable to video-recording techniques (relevant equipment out of view: n = 6;

relevant video segments missing: n = 10). The remaining ten discrepancies (38%) were felt to be attributable to the following reasons: leniency during direct observation (n = 2); leniency during video-review (n = 4); inattention during video-review (n = 3); and inattention during direct observation (n = 1).

Discussion

Comparing direct observation with video-recorded assessments, our results suggest that overall scores and pass/fail decision on the two rating tools did not significantly differ between the two modalities. Further, inter-rater reliability for both the video-recorded assessments and direct observation are acceptable and comparable. However, adequacy of video-recording may significantly impact on individual items of assessment. Thus the results of this study highlight the importance of video-recording techniques in the assessment of procedural skills.

Blinded assessments should in theory confer more objectivity than unblinded assessments.[95-97] However, to offer a truly blinded video presents a number of technical challenges to the researchers. First, audio information may potentially identify participants. On the other hand, lack of audio information, as previously noted, poses difficulty for the assessors.[99] It can be difficult to tell if erratic movements in the video are due to lack of trainee dexterity or to interruptions by faculty for other reasons.[99] Thus, audio information may be important for providing context in the interpretation of learners' movements. Potentially, a voice scrambler or voice encryption devices could be considered to anonymize audio information. Second, to ensure blinding, portions of the procedure that occur close to the participant's face or other identifying features would present a challenge for filming. In our study, in order to maintain participant confidentiality, two portions of the procedure were, in retrospect, quite difficult to capture in their entirety: wire handling and drape handling. A significant portion of

these missing segments on drape handling were felt to potentially impact assessments directly. However, if hand-positioning is adequately captured, our investigators felt that missing portions of wire handling and drape handling were not always consequential to the assessment. Thus, dynamic video recording with attention to these details (i.e. aim to capture hand-positioning rather than aim to capture the entire wire or drape) may minimize the impact of missing segments on the assessors' ability to evaluate performances.

Video-recording technique thus needs to balance between minimizing information lost with maximizing the benefit of blinding. Even if blinding is not desired or achieved, technical issues are important to consider. In another study recording scenario-based technical tasks, even with the use of pan-tilt camera and another digital camera with optical zoom, there were problems with camera angles and picture resolution.[101] Thus, attention should be paid towards key video-recording techniques such as camera angle, fields of capture, use of audio, and dynamic versus static recording. Optimization of such recording techniques may require extensive piloting.

This study is not without limitations. First, the present study only involved two assessors in a single center. Thus our study has limited ability to generalize beyond our study setting. Second, recording issues such as failure to initiate recording or forgetting to start the camera and failure to capture the end of the video due to inadequate tape capacity can potentially be minimized with the use of a research-based checklist to ensure that all steps of the research protocol are executed appropriately and in a timely fashion. Further, the use of digital cameras rather than tape-based cameras may avoid problems associated with tape capacities. However, even with the use of digital cameras, researchers need to ensure that the memory card is of sufficient capacity. Third, although based on the results of this study, a number of

recommendations could be made to improve video-recording techniques, such as the use of digital camera, voice scrambler, dynamic recording techniques, and research protocol checklists, none of these measures were tested empirically in our study. Therefore, the degree to which these recommendations will improve the validity evidence based on *response process* remains unknown. Fourth, we did not evaluate the extent to which blinding was unmasked by audio recording. Further, although we had two independent raters for most outcome measures, there was only one assessor for the qualitative evaluation of video adequacy. Lastly, the reasons for the discrepant assessment results between direct observation and video-recorded assessments were inferred retrospectively. Therefore, the degree of leniency and inattention during either modality is unknown. Future studies should consider further evaluating the relative contribution of these effects on assessments.

In conclusion, while this study did not identify significant differences in the overall checklist scores, global rating scores, or overall pass/fail decisions between direct observation and video-recorded assessments, this study did uncover a number of technical issues with respect to video-recording that merit attention. Further, these issues may directly impact upon validity evidence based on response process by impacting on data accuracy. We recommend that if video-recording is performed, consideration should be made towards using a digital camera with high storage capacity, use of dynamic recording techniques, and the use of a research protocol checklist.

CHAPTER 5: STUDY THREE: REMAINING SOURCES OF VALIDITY EVIDENCE

Adapted from "Comparing the Use of Global Rating Scale with Checklists for the Assessment of Central Venous Catheterization Skills Using Simulation"[42]

Background

In addition to the validity evidence based on content and response processes as discussed in Chapter 3 and Chapter 4 respectively, multiple additional sources of validity evidence exist. For example, additional validity evidence may be based on ***internal structure, relationship with other variables***, and ***consequences of testing***.[36]

Evidence based on ***internal structure*** refers to the degree to which the items on the test conforms to the construct of interest.[36,52] Assessing for dimensionality is one aspect of validity evidence that is based on *internal structure*.[52] For example, in the testing for competency in bedside technical skills for CVC, evidence based on *internal structure* seeks to evaluate how well do items on the assessment relate to the supposed uni-dimensionality of CVC technical skills. To do so, this study seeks to explore the dimensions captured by our constructed global rating (Appendix C), described in Chapter 4. Additional evidence based on *internal structure*[47] such as inter-rater reliability and internal consistency will also be presented.

Evidence based on ***relationship to other variables*** refers to the degree to which relationship of test scores are expected to relate to other measures.[36] For example, performance on CVC insertion is expected to improve with increasing level of training or training status. Concurrent measures of competence (checklists and global rating scales) are expected to co-vary. And perhaps even more importantly, scores on the assessment should predict clinical performance.[112] This study seeks to compare and contrast the results of assessments using

checklists and global rating scales. The rationale for comparing these assessment tools will be further described later in this section. Further, this study will evaluate additional sources of validity evidence based on *relationship with other variables* by correlating performance scores with three additional measures: 1) baseline CVC experience, 2) self-reported measures of confidence, and 3) number of attempts.

Evidence based on ***consequences of testing*** refers to the impact of the assessment on the trainees, and the sources of evidence may involve evaluating pass rates and appropriateness of cut scores set for passing.[38] To address these issues, this study seeks to compare the sensitivity and specificity of checklist scores against a measure of competence on the global rating scale.

Lastly, this chapter will revisit evidence based on ***response process***. As stated in Chapter 4, one of the more common cited sources for evidence based on *response process* stems from analyses of learners or raters' responses.[36] To better appreciate the extent to which raters' judgment on the summary global rating scale is consistent with the intended interpretation behind the rating of competent vs. not competent, this study re-evaluates all video-recorded performances rated as incompetent by the rater, but scored highly on the checklist score. Possible reasons for "failure" were evaluated.

Rationale on comparing checklists with global rating scales

While this study seeks to establish validity evidence based on the above sources of evidence, its primary focus is to compare the use of checklists with the use of global rating scales in the assessment of CVC technical competence. The rationale for this aim is as follows: as evident from our systematic-review,[40] checklists are the most commonly used assessment tool (Table 2, Chapter 3). Indeed, based on our review of 25 studies, all studies used a checklist, while only six studies used an additional global rating scale.[64-69]

The reason behind the universal and widespread adoption of the use of checklists is unknown, but is presumably related to a number of potential factors. For example, in a previously available version of the Accreditation Council for Graduate Medical Education (ACGME) toolbox of assessment methods document, checklists were considered to be the "most desirable" assessment method for assessing the performance of medical procedures, while global rating scales were rated only as "potentially applicable."[35] Second, checklists have generally been felt by raters to be more appropriate for assessments than global rating scales.[113] Third, there is a common misconception that checklists are more objective and therefore result in more reliable ratings than global rating scales.[114]

However, despite the perceived notion that checklists are more objective and reliable than global rating scales, studies in an objective structured clinical examination (OSCE) settings failed to identify higher reliability in checklists compared with global rating scales.[114,115] Further, others have found that global rating scales are in fact more sensitive than checklists in detecting increasing levels of expertise.[116,117] The reasons behind the superior psychometric properties of global ratings scales are unknown. It has been postulated that the use of checklists runs the risk of trivializing steps by rewarding thoroughness rather than clinical competence.[89,118,119]

For the assessment of technical skills, where the procedure is executed in a step-by-step fashion, on the face of it, the use of checklists makes intuitive sense. However, given existing data on the less than optimal performance of checklists in the assessments of OSCEs, empirical evidence is needed to support use of checklists in CVC assessments. Therefore, this study seeks to compare the use of global rating scale with checklists, within the context of a simulation-based formative examination on CVC insertion.

Methods

Participants

During the academic year of 2008-2009, all first year internal medicine residents were invited to participate in this study (n = 35), and only those who provided written informed consent were included in the study (n = 34; participant rate 97%). The study was approved by the university ethics review board (Ethics ID H08-01739).

Participants were enrolled in a simulator training session on CVC, as described in our previous publication.[67] Briefly, participants, in groups of 3 to 5, attended a 2-hour simulation session on CVC, taught by one faculty and one chief medical resident. Prior to training, consenting participants completed a survey on baseline central line experience and confidence (Appendix B). At the end of the simulator training session, the participants underwent a formative simulation-based assessment on their technical skills. Post assessment, participants completed a course satisfaction survey which included an item on self-reported confidence with CVC insertion (Appendix B).

As described in Chapter 4, the assessment consisted of a performance of a right internal jugular CVC insertion on a simulator (Laerdal IV Torso; Laerdal Medical Corp, Wappingers Falls, New York), using a standard kit provided (Arrow® multi-lumen central venous catheterization kit with Blue FlexTip®, Reading, PA). Participants were instructed to perform the CVC as they would in real-life, without externally imposed time limits on the procedure. Feedback about their performances was given only at the end of the procedure. Participants who failed the formative examination were requested to enroll in additional practice sessions. Each examination performance was video-recorded in a blinded fashion, as described in detail in Chapter 4.

Assessment Tools

Three tools are used for this study for the video-recorded performances: a global rating scale, a 10-item checklist, and a 21-item checklist. All tools were adapted such that items deemed not applicable to our simulation-based assessments were removed. Rather than assuming validity of these modified tools, content validity of each of the three tools was re-addressed through input from an expert panel, consisting of one nephrologist, two internists, one intensivist, and one general surgeon. Consensus was achieved with the final items.

Global Rating Scale

The 8-item global rating scale (Appendix C) has been described in detail in Chapter 4. This 8-item scale has a summary item, rated on a 6-point Likert scale, with descriptive anchors ranging from 1 = "not competent to perform independently" to 6 = "above average competence to perform independently."

Pass/fail decisions are made by dichotomizing the scores for this summary item such that a score of 3 or more was considered "competent to perform the procedure," while a score or 2 or less was considered "not competent to perform the procedure."

10-item Checklist

The 10-item binary checklist (Appendix D) has also been described in detail in Chapter 4. Briefly, this checklist was constructed based on cognitive task analysis and expert input.[55] Items on this tool mapped to 6 out of the 7 key procedural domains outlined in Chapter 3. Only the domain of "communication and working with the patient," which was not deemed applicable in this current examination, was not assessed by this tool (Table 3).

Raters using this tool were also asked to record the number of attempts made by the participants at locating the vein.

21-item Checklist

The 21-item binary checklist (Appendix E) is adapted from a previously published 27-item checklist.[72] This tool was initially constructed based on literature review, expert panel input, as well as a rigorous checklist construction process.[72] All seven procedural domains are assessed by this checklist (Table 3).

For both of these checklists, the overall checklist score was calculated as a number of completed items divided by the total number of items, presented as percentage.

Expert Raters

Two expert raters (IM, NZ) independently evaluated all 34 video-recorded performances. Raters were both faculty members with a minimum of three years of experience in simulation-based CVC teaching and assessment and ten years of clinical experience in performing CVCs. As noted in the preceding chapter, each rater was trained for a minimum of three hours on the use of each of the assessment tools, by review of video-recorded performances that represented a range of performances from poor to excellent. Post-training, the ICCs of the raters' assessment scores were > 0.80.

Video-recorded performance assessment procedure

Video-recorded performances from miniDV formats were reformatted into DVDs for distribution to expert raters. To minimize the extent to which one tool may have systematically influenced the rating of the subsequent tool, for each rater, 50% of the videos were rated first using the 10-item checklist while the remaining 50% of the videos were rated first using the global rating scale. The 21-item rating tool was used independently by the raters on the same videos approximately one year subsequent to the initial video-ratings, as the original tool had not been published at the time of initial video reviews.[72] Given the delayed time frame and the

random order in which the DVDs were reviewed, it was felt that the rating using the 21-item tool was not subject to the influence of rating by another tool.

Estimation of the risk of one tool influencing the rating of the subsequent tool

As two tools were used during the initial video review, despite altering the order in which the tool was used, there remains a possibility that one tool may have systematically influenced the rating of the subsequent tool. To estimate the likelihood of the presence of this issue, the same two raters, two years after the completion of the study, re-analyzed each video. During this re-analysis, all assessment tools that were initially rated *after* the completion of another tool were re-analyzed independently. This re-analysis allows for the assessment of reliability of ratings between tools completed with and those completed without the influence of another tool. Average ICC for the global rating score between ratings with and without the influence of another tool was 0.92. Overall Kappa score for the 10-item checklist was 0.92, with a summary checklist score reliability of 0.97. As stated previously, this process was not conducted for the 21-item tool as this tool was rated independently without the use of another tool.

Relationship to other variables

Relationship of checklist scores to global rating scale scores was assessed. In addition, performance score measures with three additional variables were evaluated: 1) baseline reported number of internal jugular CVCs inserted, 2) self-reported confidence post-training, rated from 0 to 5, where 0 = no confidence and 5 = complete confidence, and 3) observed number of attempts to locate the vein.

Statistical analysis

To explore the dimensions assessed by the global rating scale, the following analyses were performed: after confirming the appropriateness of performing principal component analysis using Bartlett's test of sphericity (Chi-square = 114.7, $p < 0.001$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (0.60), principal component analysis was performed on the eight items in the global rating scale, using a VARIMAX rotation. A Scree plot was inspected.[120] Factors with eigenvalues greater than 1 were retained. Item loadings < 0.40 are suppressed in the publication,[42] but reported in full in this chapter.

Internal consistency was evaluated using Cronbach's alpha (α). Inter-rater reliability was evaluated using ICCs, Pearson's correlation coefficient, Kendall's tau-b coefficient, and Cohen's Kappa where appropriate. Both correlation coefficients and disattenuated coefficients are reported.[121] Disattenuated coefficients represent the hypothetical correlation between two measures assuming the two measures are perfectly reliable.

The sensitivity and specificity of the overall scores on the checklists against the dichotomous measure of competence on the global rating scale (a score of 3 or more) were evaluated at various checklist cutpoints with a Receiver Operating Characteristic (ROC) analysis. The area under the curve (AUC) was then estimated and used as an index of diagnostic accuracy.

Comparisons between groups were made using Student's *t* tests, Chi-square, and Wilcoxon rank-sum tests where appropriate. All analyses were performed using PASW Statistics software, version 18.0 for Windows (PASW, IBM Corporation, Somers, NY) and Stata 11.0 (StataCorp LP, College Station, TX).

Results

Of the 34 participants who completed the study protocol, 21 (62%) were deemed competent at inserting CVC independently, while 13 (38%) were deemed not to be competent to do so. Baseline demographics of the participants who were and were not deemed competent did not differ (Table 8).

Table 8. Demographic characteristics of the 34 participants

Characteristic	**Deemed Competent to Perform Independently or Better (n=21)**	**Deemed Borderline or Not Competent to Perform Independently (N=13)**	***p* - value**
Gender – no. (%)			
Male	14 (67)	8 (62)	0.76
Female	7 (33)	5 (38)	
Completed number of months of training in intensive care unit – no. (%)			
None	9 (43)	6 (46)	0.91
One	11 (52)	6 (46)	
Two	1 (5)	1 (8)	
Number of central venous catheterization inserted by others observed – median (IQR*)			
Femoral	2 (1-5)	2 (1-5)	0.55
Internal jugular	2 (1-10)	4 (1-5)	0.57
Subclavian	2 (0-2)	1 (0-2)	0.66
Number of central venous catheterization attempted or performed – median (IQR*)			
Femoral	2 (0-2)	1 (0-5)	0.78
Internal jugular	3 (1-6)	2 (0-4)	0.32
Subclavian	0 (0-1)	0 (0-2)	0.34

* IQR denotes interquartile range.

Validity evidence based on internal structure: dimensions assessed

For assessing the global rating scale's dimensionality, we identified two factors with eigenvalues greater than 1, accounting for 84.1% of the overall variance. Post-rotation, five

global rating scale items loaded on the first factor, while two factors loaded on the second (Table 9). The first factor consisted primarily of behaviours that can be characterized as relating to "technical ability" ($\alpha = 0.78$). The second factor consisted of behaviours that relate to procedural "safety" ($\alpha = 0.76$). The item "appropriate preparation of instruments pre-procedure" did not load on any factor.

Table 9. Rotated factor loadings for scale items.

Global Rating Scale Item	Factor 1 - Technical Ability	Factor 2 - Safety
Knowledge of instruments	0.93*	0.09
Instrument handling	0.90*	-0.07
Flow of procedure and forward planning	0.90*	0.26
Time and motion	0.88*	0.06
Appropriate analgesia	0.47*	-0.40
Seeks help where appropriate	0.25	0.87*
Aseptic technique	-0.07	0.59*
Appropriate preparation of instruments pre-procedure	-0.06	-0.18

* Factor loadings ≥ 0.40

Correlation of the checklist scores with factor scores on the global rating scale

Inter-rater reliability of ratings from the global rating scale and the two checklists is shown in Table 10.

The correlation between the overall 10-item checklist score and the weighted factor score on ***technical ability*** was positive (0.49; 95% confidence interval 0.17 to 0.71), while the correlation between the 10-item checklist score and the weighted factor score on ***procedural safety*** was negative (-0.17; 95% confidence interval -0.48 to 0.18).

Table 10. Inter-rater reliability (intraclass correlation coefficients or Kappa statistics) for items in the global rating scale and the two checklists

Global Rating Scale Items (α = 0.79)*	**Intraclass correlation coefficient**	**95% Confidence Interval**
Appropriate preparation of instruments pre-procedure	-0.04	-0.38 to 0.30
Appropriate analgesia	0.67	0.42 to 0.82
Time and motion	0.52	0.23 to 0.73
Instrument handling	0.63	0.33 to 0.80
Flow of procedure and forward planning	0.30	-0.03 to 0.57
Knowledge of instruments	0.44	0.12 to 0.67
Aseptic technique	0.62	0.36 to 0.79
Seeks help where appropriate	0.63	0.37 to 0.80
Overall ability to perform procedure	0.63	0.38 to 0.80
Overall agreement on competence	Kappa = 0.52	0.23 to 0.81
Ten-Item Checklist (α = 0.67)	**Kappa Statistic**	**95% Confidence Interval**
Prepared site appropriately	0.77	0.46 to 1.00
Identified proper landmarks	Perfect agreement	
Inserted needle at correct angle	0.39	0.00 to 0.78
Aspirated while inserting needle	*NA (no variability in rater 2)	
Inserted guidewire appropriately	0.65	0.03 to 1.00
Withdrew needle and incised skin	0.79	0.38 to 1.00
Inserted catheter and withdrew wire	Perfect agreement	
Aspirated blood and flushed ports	0.64	0.18 to 1.00
Occluded ports	0.62	0.24 to 1.00
Secured catheter	0.84	0.54 to 1.00
Inter-rater reliability of overall ten-item checklist score	$r = 0.88$; $p < 0.001$	
Twenty-one Item Checklist (α = 75)		
Flush the ports on the catheter with sterile saline	Perfect agreement	
Clamp each port	Perfect agreement	

Remove brown port from end of catheter to accommodate wire	0.18	0.00-0.58
Area is cleaned with chlorhexidine	Perfect agreement	
Resident gets in sterile gown, gloves, hat and mask	Perfect agreement	
Area is draped in usual sterile fashion (*must be full body drape)*	0.47	0.00-1.00
The vein is localized using anatomical landmarks	Perfect agreement	
The skin is anesthetized with 1% lidocaine in a small wheal	0.84	0.54-1.00
The deeper structures are anesthetized	0.89	0.68-1.00
Using the large needle or catheter-syringe complex, cannulate the vein while aspirating	Perfect agreement	
Remove the syringe from the needle or advance the catheter into the vein removing both the syringe and the needle	Perfect agreement	
Advance the guidewire into the vein no more than approximately 12-15cm	0.53	0.14-0.93
Knick the skin with the scalpel to advance the dilator	0.79	0.38-1.00
Advance the dilator over the guidewire and dilate the vein	Perfect agreement	
Advance the triple lumen over the guidewire	0.65	0.02-1.00
Never let go of the guidewire	0.64	0.35-0.93
Once the catheter is inserted remove the guidewire in its entirety	Perfect agreement	
Advance the catheter to approx to 14-16 cm on the right side	0.74	0.47-1.00
Ensure there is blood flow/flush each port	0.78	0.38-1.00
Secure the catheter in place	Perfect agreement	
Maintain sterile technique	0.62	0.22-1.00

Inter-rater reliability of overall 21-item checklist score	$r = 0.89$; $p < 0.001$	

* α refers to Cronbach's alpha, an assessment of internal reliability.

The correlation between the overall 21-item checklist score and the weighted factor score on ***technical ability*** was also positive (0.43; 95% confidence interval 0.10 to 0.67). The correlation between the overall 21-item checklist score and the weighted factor score on ***procedural safety*** was also negative (-0.13; 95% confidence interval -0.45 to 0.22).

Validity evidence based on relationship with CVC experience, confidence, and number of attempts

At baseline, participants reported minimal CVC experience: median number of internal jugular CVCs placed: 3, IQR 0-5. Prior to training, participants reported low confidence (mean score 2.2 ± 0.22, where 0 = no confidence, and 5 = complete confidence). Post-training, there was a statistically significant increase in confidence (mean score post-training 3.5 ± 0.75, $p < 0.001$). Inter-rater reliability for the number of attempts to locate the vein was high (0.89, 95% CI 0.79-0.94).

Baseline experience pre-training did not correlate with performance scores (Table 11). Self-reported confidence post-training positively correlated with the 10-item checklist score, overall global rating scale score, and the weighted factor 1 score (technical ability). Number of attempts to locate the vein correlated negatively with the 10-item checklist score, 21-item checklist score, and the weighted factor 1 score on ***technical ability*** (Table 11).

Table 11. Performance score' relationship with baseline central venous catheter experience, confidence, and number of attempts

Performance scores	**Pearson's correlations with baseline number of IJ CVC*s performed (*r*, *p*-value)**	**Kendall's tau-b for correlations with self-reported confidence post-training† (τ, *p*-value)**	**Pearson's correlations with observed number of attempts to locate the vein (*r*, *p*-value)**
10-item Checklist Score	-0.28, $p = 0.10$	0.44, $p = 0.005$‡	-0.37, $p = 0.01$‡
21-item Checklist Score	-0.21, $p = 0.23$	0.18, $p = 0.23$	-0.35, $p = 0.04$‡
Overall Global Rating Scale Score	0.13, $p = 0.47$	0.36, $p = 0.03$‡	-0.27, $p = 0.12$
Weighted Factor 1 Score	0.34, $p = 0.051$	0.43, $p = 0.003$‡	-0.33, $p = 0.022$‡
Weighted Factor 2 Score	0.26, $p = 0.13$	-0.03, $p = 0.86$	0.073, $p = 0.68$

* IJ CVCs refers to internal jugular central venous catheterizations.
† Confidence was rated from 0 to 5, where 0 = no confidence, and 5 = complete confidence.
‡ Denotes $p < 0.05$

Checklists versus global rating scale

Based on the expert global judgment of competence, 21 participants (62%) were rated overall as being competent, while 13 (38%) were rated as not competent to independently place a CVC.

The mean overall score on the 10-item checklist for those deemed competent was 95.2 ± 8.1%, which is significantly higher compared to those who were deemed not competent (81.0 ± 18.6%, $p = 0.002$). Correlation between the overall 10-item checklist score with the summary measure on the global rating scale was high ($r = 0.58$, $p = 0.0003$). Corrected for attenuation, the correlation was 0.80.

Using the 21-item checklist, the mean overall score for those deemed competent was not significantly different from those deemed not competent (92.0 ± 5.2% vs. 84.1 ± 14.7% respectively, $p = 0.08$). The correlation between the overall 21-item checklist score with the summary measure on the global rating scale was high ($r = 0.60$, $p = 0.0002$). Corrected for attenuation, the correlation was 0.79.

Validity evidence based on consequences of testing

On ROC analyses, the overall 10-item checklist score had an acceptable discrimination (AUC = 0.79, standard error = 0.078, 95% confidence interval 0.64 to 0.94),[122] while the 21-item checklist's AUC was 0.68, standard error = 0.098, 95% confidence interval 0.48-0.87). Table 12 shows the sensitivity and specificity for different cutpoints for the checklist score. For maximum sensitivity (100%), a cutpoint of 80% was chosen as the optimal cutpoint for both checklists. At this threshold, a sensitivity of 100% allows us to reliably "rule out" competence for individuals with a checklist score of < 80%. However, the poor specificity for both checklists did not allow us to "rule in" competence despite high checklist scores.

Table 12. Sensitivity and specificity of various cutpoints on the overall checklist scores in determining competence, as diagnosed based on global rating scale.

Cutpoint for Passing	Sensitivity for 10-item Checklist	Sensitivity for 21-item Checklist	Specificity for 10-item Checklist	Specificity for 21-item Checklist
≥ 30%	100%	100%	0%	0%
≥ 60%	100%	100%	7.69%	7.69%
≥ 80%	100%	100%	23.08%	23.08%
≥ 90%	80.95%	76.19%	53.85%	53.85%
100%	28.57%	9.52%	15.38%	100%

Validity evidence based on response process: analysis of raters' assessments

Out of the 13 participants who were deemed, based on the global rating scale, incompetent in being able to insert CVCs independently, 11 individuals scored ≥ 80% on the checklists. Retrospective re-analysis of videos revealed the following reasons for incompetence in these 11 individuals: significant breaches in sterility (n = 5), loss of wire control (n = 4; median duration of time without wire control: 35 seconds, range 17-38 seconds), multiple attempts (n = 2; both cases > 6 attempts were made).

Discussion

For the assessment of technical competence of CVC insertion in a formative simulation-based examination, this chapter presented validity evidence based on *internal structure*, *relationship with other variables, consequences of testing*, and *response process*.

Comparing the use of three assessment tools: a global rating scale and two checklists, our results suggest that for the assessment of competency in CVC skills, the use of checklists is not always the "most desirable" evaluation method, as the ACGME had previously endorsed.[35]

With respect to validity evidence based on *internal structure*, our results indicate that two dimensions were captured by the global rating scale: ***technical ability*** and ***procedural safety***. Neither checklist adequately captured errors relating to safety issues. In fact, checklists scores were uniformly negatively associated with factor scores on ***procedural safety***, although the association was not statistically significant.

This finding is consistent with the literature review on procedural checklists in general, where 30-50% of checklists did not assess for competencies in the area of "infection control" and "safety."[60] Specific to the CVC, this finding is also consistent with our systematic review on

CVC checklists, as reported in Chapter 3, where 16% of checklists did not assess for "infection control" and 28% of checklists did not assess for "safety."[40]

With respect to validity evidence based on *consequences of testing*, our raters deemed a number of performances, on global rating scale, as being incompetent to perform CVC independently. Out of these individuals, the majority scored highly on the checklists. In order to assess the potential *consequences to testing*, we gathered further validity evidence based on *response process* by reanalyzing raters' assessments. Indeed, based on further review, all those who were deemed incompetent committed errors that were considered serious in nature. In particular, breaches in sterility, loss of wire control, and an unsafe number of attempts at venous access are all errors associated with patient safety implications. Infectious complications relating to CVC can be as high as 26%.[5] Therefore, meticulous attention to sterility is an important aspect of the procedure. Loss of wire control likewise has significant safety implications. In a study evaluating malpractice claims for CVC, the most common complication was wire/catheter embolism.[123] Lastly, the incidence of mechanical complications increases significantly with 3 or more attempts at insertion.[5,124] Therefore, multiple attempts by a trainee should be flagged as problematic.

While the commission of the aforementioned serious errors appeared to have resulted in an overall global impression of incompetence, commission of these same errors resulted in only a minimal reduction in the checklist scores. Our study identified a positive correlation of both checklists scores with the technical ability factor score on the global rating scale. However, as previously stated, we found a negative correlation of both checklist scores with the safety factor score. Although differences in the two correlations were not statistically significant, the two

differed in the direction, thereby lending support to the impression that these checklists perhaps capture items relating to technical ability more than they do safety parameters.

Overall, our results indicate that the use of checklist scores in diagnosing competence was associated with a higher sensitivity than specificity. A low checklist score (<80%) was uniformly associated with procedural incompetence, while a number of individuals with high checklist scores (≥80%) were still deemed incompetent. All of these candidates committed errors that were considered serious in nature. That is, use of a high checklist score alone to diagnose competence may result in a number of false positives.

What are the implications of these results? Consistent with the literature on OSCE assessments, our results suggest that checklists should not automatically be assumed to be the preferred method of assessment. All pass/fail decisions require value judgments.[39] Setting lower cut scores may be appropriate for high-stakes examinations, where the misdiagnosis of incompetence in an otherwise competent individual may have devastating consequences. However, in a formative assessment setting, educators may instead wish to minimize the number of incompetent trainees missed by the assessment by either using a higher cut score, or alternatively, using a global rating scale instead. As our study results indicate, high cut scores may still miss incompetent performances. Thus, the use of global ratings scale should be considered. While we do not advocate that checklists be abandoned altogether for the assessment of procedural skills, we do however recommend that validity evidence for their use be evaluated prior to their use.

Both of the checklists evaluated in our study were constructed carefully; one used a cognitive task analysis approach,[55] while the other used a rigorous checklist development procedure.[72] Nonetheless, despite careful construction, improvement in content validation may

be made to these tools either by including additional items of safety parameters or additional clinically discriminating items.[125-127]

Limitations

This study has several important limitations, including the fact that the study was performed in a single centre with a relatively small sample size. Secondly, in the absence of an independent gold standard measure, it may be problematic to use physicians' subjective judgment on the global rating scale as the standard against which checklist scores were compared. One can also argue for the use of checklist scores as the standard instead. However, our re-analysis of video-reviews confirmed the legitimacy of the global rating judgments. Thus, the use of global rating scale in this setting likely is justified. Further, our study was performed in the setting of a formative examination. That is, we wished to maximize the number of trainees identified as "incompetent" and who may benefit from additional practice. We were willing to accept some degree of false positives in the identification of incompetence. However, we were less willing to accept false negatives (i.e., missing individuals who may need additional instruction or practice). Indeed, all candidates deemed incompetent based on a low checklist score were also deemed incompetent by the global rating scale, while the use of global rating scale identified additional incompetent performances that were rated highly by both checklists.

Thirdly, the results of our study conclusions regarding the use of assessment tools are context-specific. For example, our conclusion relate to the use of the two checklists and the global rating scale used in our study, in a formative simulation-based examination on CVC performance by first year medical residents, using landmark technique, evaluated by expert trained raters. Checklists constructed in a different manner may outperform our global rating scale. Likewise, the reliability of scores from these assessment tools is unknown in the hands of

untrained raters or on CVC performances on patients. Context, therefore, needs to be taken into account in the interpretation of our results.

Fourthly, although we attempted to estimate the degree to which the use of one tool influenced the next, our raters were trained on the use of both checklists and global rating scale. Therefore, we cannot exclude the possibility that intimate knowledge of items on the checklists may have influenced assessments using the global rating scale, and vice versa, even when the tools were not filled out at the same time.

Fifthly, we did not explore the effects of modifying checklists, such as differential weighting of critical items, or the use of three-point scale for each item on the checklists rather than their current binary form. While such alternative scoring schemes are appealing in theory, their use had not been previously identified to result in superior rating to traditional scoring methods.[128,129] Nonetheless, the incorporating of clinically discriminating items showed promise in the OSCE setting[126,127] and should be further evaluated.

Finally, the validity evidence of scores from our constructed global rating scale cannot be assumed, despite the fact that it was constructed based on two previously validated tools.[109] The compilation of two tools into one resulted in the inclusion of behaviorally anchored scales for some items but not for others. The uneven distribution of anchors may have resulted in the variable inter-rater reliability observed amongst items on the global rating scale as well as the differential weighting on the factor scales. Furthermore, the categories for competence was chosen arbitrarily, albeit by consensus, based on concerns that assessments may be positively skewed.[59] As a result, three categories were available to assist examiners in differentiating among above average performances. The trained raters in our study, however, ultimately deemed 38% of the candidates as incompetent at placing a CVC independently. Therefore, in future

studies, consideration should be made for the inclusion of additional categories to assist examiners in differentiating among below average performances rather than above average performances.

Conclusion

Despite these limitations, results from our study raise an important question regarding the practice of automatically adopting checklists as the preferred method of assessment of procedural skills. Our study provides an example whereby the use of a global rating scale may in fact be preferred over the use of either of the two checklists in our study. Future study should focus on further optimizing the construction of assessment tools and correlating assessment results with clinical outcomes.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

Summary of Findings

By way of the three aforementioned publications,[40-42] this work seeks to accomplish the following two goals: first, based on the unified theory of validity,[36] this work presents evidence of validity from each source of validity evidence for the assessment of technical competence in medical trainees using a formative simulation-based examination: ***content, response process, internal structure, relationship with other variables, and consequences of testing.***[36] The intended interpretation of scores derived from this assessment was to identify junior trainees who would require additional training. Our studies presented robust but imperfect evidence for each of the above sources of validity. Based on a systematic review of the literature, study one has provided the readers with a comprehensive review of the underlying task of CVC insertion.[40] From this, two checklists and a constructed global rating scale were further evaluated in study three.[42] Items on both checklists demonstrated validity evidence based on content and map to at least six of the seven main procedural competency domains.[60] Evidence based on internal structure suggests that the dimensionality of the constructed global rating scale assessed ***technical ability*** and ***procedural safety***.[42] Despite validity evidence demonstrated by the two chosen checklists, overall, our results suggest that, in the hands of expert raters, the use of the global rating scale was superior at identifying individuals in need of further training, compared to the use of the two checklists. Thus, as a consequence of testing, the use of checklists alone may miss individuals whose performances demonstrated serious errors. Lastly, assessments by video-recorded reviews demonstrated high correlations with assessments performed by direct observation.[41] However, educators need to be mindful of technical issues and video recording

techniques that may impact on the individual items of assessments, and affect the accuracy of the assessment data.

Limitation of program of research

In addition to the limitations of individual studies, as highlighted in each chapter, this program of research has a number of limitations that deserve mention. First, this work pertains to the assessment of technical skills only. We chose to limit our focus on technical skills competence which we felt was already challenging enough to properly define and assess, as potentially reliable assessments may be limited to one to two dimensions only.[130,131] However, insertion of CVC is inherently a holistic skill that requires additional competence in technical as well as non-technical domains such as communication skills, professionalism, and team work. Appropriate assessment of these additional domains requires additional validity evidence and is beyond the scope of this work. We refer the readers to others who have done work in these domains.[60,88,132]

Second, the primary purpose of the assessment described in this work deserves highlighting: to identify junior medical residents who may benefit from additional training before being able to perform the procedure independently. We do not, at this time, have performance data to support that notion that these identified individuals ultimately benefitted from additional training. However, work by others suggest that additional training until sufficient mastery level likely is associated with improved patient outcomes.[11] We do not have any reasons to believe that our learners who were deemed not competent would not similarly benefit from further training. Further, as a corollary to our stated primary objective, it is worth highlighting that our work does not have performance data to support the notion that those who "passed" this

assessment process can indeed proceed to place a CVC safely and successfully without supervision. Additional predictive validity data are needed to support this claim.

Third, our focus was on identifying competence and ***not*** technical expertise. True expertise requires extensive deliberate practice on the part of the proceduralist and is beyond the scope of this work.[19,20,133] However, despite the mandate that CVC insertion is a mandatory skill for internal medicine,[134] the number of CVCs that the internists place each year is only an estimated median of five per year.[135] Therefore, this low number is unlikely to render the average internist a true expert in this skill or allow one to maintain expertise. Nonetheless, while expertise is a lofty goal worthy of pursuit, in the spirit of the competency-based physician framework, the critical goal for the average internist is to be able to first and foremost ***safely and successfully*** perform the procedure for our patients on those who require one from us.

Fourth, this work does not specifically address instructional design and the importance of mastery learning, which has been previously demonstrated in a number of settings.[11,21,71,137,138] Although the assessment process described in this work was part of the instructional design, by necessity, mastery-based learning is an iterative process. Therefore, the educational module followed by an assessment, as described in this work, is only a small part of the overall educational design. Further, regardless of any of the presented validity evidence, a one-time assessment procedure is only one step within the larger framework of a programme of assessment.[139,140] Comprehensive assessment of competence in CVC *as a whole* requires an application of a programmatic approach.[36,140] Additional approaches to assessments should be considered, such as incorporating assessments of varying difficulty level and modalities,[141] incorporating work-place based assessments, ensuring fairness in testing for all trainees,[36]

evaluating for the educational impact of the assessments, and assessing for feasibility and acceptability.[142]

Fifth, we did not address the role of the learner or the potential impact on learning and assessments by learner factors such as motivation, attitude, inattention, or fatigue.[143,144] Sixth, our assessment process for CVC did not incorporate the use of ultrasound. The use of ultrasound for CVC guidance is considered the standard of care[34,145] and therefore its use should be taught and assessed. In a subsequent study that is currently under consideration by a journal, we have evaluated further the instructional methods and assessments for ultrasound-guided CVC.

Last but not least, despite our study's attempt to address each source of validity evidence, it is worth re-emphasizing that we have not addressed all possible sources of validity evidence. Ultimately, a validity argument will need to integrate various sources of evidence into a coherent whole.[39,44,45] Overall, while this work present likely sufficient evidence to support the use of the assessment procedure described for the identification of learners in need of further training, further evidence is required to justify its use beyond this application.

Future Directions

In many ways, the validation process may appear to never end, at least until the point at which the validity evidence for the intended use appears reasonably well-supported and defensible.[36] While future studies that should be performed are many, based on this work, some proximal questions should be addressed.

First, do our results regarding checklists versus global rating scales generalize to the assessment of other skills? To that end, we have now additional data to support the notion that checklists have low specificity in the diagnosis of competence in six additional bedside procedural skills.[146] That is, high checklist scores did not preclude the diagnosis of incompetence

in other skills such as lumbar puncture, thoracentesis, paracentesis, arterial blood gas sampling, intubation, and knee arthrocentesis, suggesting that this may be a consistent pattern. However, whether this phenomenon occurs in other centers should be confirmed.

Second, can the incorporation of clinically-discriminating items in procedural checklists improve their diagnostic ability for procedural skills competence?[125] We are currently investigating the ability for expert-consensus based procedural error items to identify procedural competence/incompetence. If further modifications of checklists are unable to improve their diagnostic ability, then the question remains, how do global rating scales perform alone, and how do they perform in the hands of non-experts? Much additional work remains to be done to further support a procedural assessment process that yields valid and reliable results.

Conclusions

Based on the unified theory of validity,[36] this work presents the results of three studies, with an overarching goal to examine various sources of validity evidence. This work seeks to support the interpretation of scores for the assessment of technical competence in medical trainees using a formative simulation-based examination on CVC. Specifically, as a systematic review of assessment tools in the literature, study one (Chapter 3) presented ***validity evidence based on content***, to better define the construct of CVC insertion.[40]

Study two (Chapter 4) presented sources of evidence based on one aspect of ***response process*** that relates to rating data accuracy.[41] Specifically, although compared to assessments based on direct observation, video-recorded assessments did not differ in terms of overall scores or pass/fail decisions, specific aspects of video-recording techniques do impact on individual items within the assessment tools.[41] In order to maintain blinding of the participants, our study identified that wire and drape handling were difficult to capture on video-recording. Therefore,

we recommend additional video-recording techniques such as maximizing camera angle, fields of capture, and use of dynamic video recording in future studies.

Study three (Chapter 5) presented sources of validity evidence based on ***internal structure, relationship with other variables, and consequences of testing.***[42] In this study, as evidence based on internal structure, our global rating scale assessed two predominant domains: technical ability and procedural safety. Further, we identified that although high scores on the checklists were able to identify competence performances, they did not preclude incompetence. That is, the use of checklists yielded false negatives with respect to the identification of incompetence. As such, for formative assessments where false negatives are important to avoid, we recommend the use of global rating scales.

CVC is a technically challenging procedure that is important for patient care clinically. The need to assess for procedural competence in a valid and reliable manner is important for patient safety reasons. The results of this thesis serve to provide validity evidence on the assessment process: how to assess for procedural competence and what tools to use. Importantly, they provide evidence for identifying trainees who may benefit from further training. On a more practical basis, the assessment process described in this thesis helps identify individuals who, if given the opportunity to perform the procedure clinically, should do so only under very close supervision. Alternatively, an argument could be made that these trainees should not perform the procedural at all clinically until further competency is attained. Future studies should seek to evaluate the validity of this argument.

References

1. Momiy J, Vasquez J. Iatrogenic vertebral artery pseudoaneurysm due to central venous catheterization. Proceedings (Baylor University Medical Center) 2011;24:96.

2. Aubaniac R. Subclavian intravenous injection; advantages and technic. La Presse Medicale 1952;60:1456.

3. Mermel LA. Prevention of intravascular catheter–related infections. (Erratum: Ann Intern Med 133:395, 2000). Ann Intern Med 2000;132:391-402.

4. Centers for Disease Control and Prevention. Morbidity and Mortality Weekly Report. Vital signs: central line--associated blood stream infections --United States, 2001, 2008, and 2009. (Accessed October 3, 2014, at http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6008a4.htm.)

5. McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med 2003;348:1123-33.

6. Eisen LA, Narasimhan M, Berger JS, Mayo PH, Rosen MJ, Schneider RF. Mechanical complications of central venous catheters. J Intensiv Care Med 2006;21:40-6.

7. Warren DK, Quadir WW, Hollenbeak CS, Elward AM, Cox MJ, Fraser VJ. Attributable cost of catheter-associated bloodstream infections among intensive care patients in a nonteaching hospital. Crit Care Med 2006;34:2084-9.

8. Blot SI, Depuydt P, Annemans L, et al. Clinical and economic outcomes in critically ill patients with nosocomial catheter-related bloodstream infections. Clin Infec Dis 2005;41:1591-8.

9. Pronovost P, Needham D, Berenholtz S, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. N Engl J Med 2006;355:2725-32.

10. Berenholtz SM, Pronovost PJ, Lipsett PA, et al. Eliminating catheter-related bloodstream infections in the intensive care unit. Crit Care Med 2004;32:2014-20.

11. Barsuk JH, Cohen ER, Feinglass J, McGaghie WC, Wayne DB. Use of simulation-based education to reduce catheter-related bloodstream infections. Arch Intern Med 2009;169:1420-3.

12. Sherertz RJ, Ely EW, Westbrook DM, et al. Education of physicians-in-training can decrease the risk for vascular catheter infection. Ann Intern Med 2000;132:641-8.

13. Warren DK, Zack JE, Cox MJ, Cohen MM, Fraser VJ. An educational intervention to prevent catheter-associated bloodstream infections in a nonteaching, community medical center. Crit Care Med 2003;31:1959-63.

14. Warren DK, Zack JE, Mayfield JL, et al. The effect of an education program on the incidence of central venous catheter-associated bloodstream infection in a medical ICU. Chest 2004;126:1612-8.

15. Graham AS, Ozment C, Tegtmeyer K, Lai S, Braner DAV. Central Venous Catheterization. N Engl J Med 2007;356:e21-.

16. Armstrong CW, Mayhall CG, Miller KB, et al. Prospective study of catheter replacement and other risk factors for infection of hyperalimentation catheters. J Infect Dis 1986;154:808-16.

17. Sznajder JI, Zveibil FR, Bitterman H, Weiner P, Bursztein S. Central vein catheterization. Failure and complication rates by three percutaneous approaches. Arch Intern Med 1986;146:259-61.

18. Fares LG, 2nd, Block PH, Feldman SD. Improved house staff results with subclavian cannulation. Am Surg 1986;52:108-11.

19. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79:S70-81.

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

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

22. Ma IW, Brindle M, Ronksley P, Lorenzetti D, Sauve R, Ghali W. Use of simulation-based education to improve outcomes of central venous catheterization: a systematic review and meta-analysis. Acad Med 2011;86:1137-47.

23. Snell LS, Frank JR. Competencies, the tea bag model, and the end of time. Med Teach 2010;32:629-30.

24. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach 2010;32:638-45.

25. Iobst WF, Sherbino J, Cate OT, et al. Competency-based medical education in postgraduate medical education. Med Teach 2010;32:651-6.

26. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010;32:676-82.

27. Block JH, Burns RB. Mastery learning. Review of Research in Education 1976;4:3-49.

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

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

30. Barsuk JH, Cohen ER, Potts S, et al. Dissemination of a simulation-based mastery learning intervention reduces central line-associated bloodstream infections. BMJ Quality & Safety 2014;23:749-56.

31. Zendejas B, Cook DA, Bingener J, et al. Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized controlled trial. Ann Surg 2011;254:502-9.

32. O'Grady NP, Alexander M, Burns LA, et al. Guidelines for the prevention of intravascular catheter-related infections. Clin Infect Dis 2011;52:e162-93.

33. Moureau N, Lamperti M, Kelly LJ, et al. Evidence-based consensus on the insertion of central venous access devices: definition of minimal requirements for training. Br J Anaesth 2013;110:347-56.

34. National Institute for Health and Clinical Excellence (NICE). Technology Appraisal No 49: Guidance on the use of ultrasound locating devices for placing central venous catheters. (Accessed October 9, 2014, at http://www.nice.org.uk/Guidance/TA49/Guidance/pdf/English.)

35. ACGME Competencies: Suggested Best Methods for Evaluation. Version 1.1. The Accreditation Council for Graduate Medical Education. (Accessed November 22, 2010, at http://www.acgme.org/Outcome/assess/ToolTable.pdf.)

36. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.

37. Messick S. Standards of validity and the validity of standards in performance asessment. Educational Measurement: Issues and Practice 1995;14:5-8.

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

39. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for educational and psychological testing. Washington, D.C.: American Educational Research Association; 1999.

40. Ma I, Sharma N, Brindle M, Caird J, McLaughlin K. Measuring competence in central venous catheterization: a systematic-review. SpringerPlus 2014;3:33.

41. Ma IWY, Zalunardo N, Brindle ME, Hatala R, McLaughlin K. Notes from the Field: Direct Observation Versus Rating by Videos for the Assessment of Central Venous Catheterization Skills. Evaluation & the Health Professions 2014.

42. Ma IW, Zalunardo N, Pachev G, et al. Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. Adv Health Sci Educ Theory Pract 2012;17:457-70.

43. Kane MT. Current concerns in validity theory. Journal of Educational Measurement 2001;38:319-42.

44. Cook DA. When I say… validity. Med Educ 2014;48:948-9.

45. Kane MT. An argument-based approach to validity. Psychological Bulletin 1992;112:527.

46. Guion RM. On trinitarian doctrines of validity. Professional Psychology 1980;11:385-98.

47. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Advances in Health Sciences Education 2014;19:233-50.

48. Haynes SN, Richard D, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychological assessment 1995;7:238.

49. Sireci S, Faulkner-Bond M. Validity evidence based on test content. Psicothema 2014;26:100-7.

50. Fitzpatrick AR. The meaning of content validity. Applied Psychological Measurement 1983;7:3-13.

51. Padilla J-L, Benítez I. Validity evidence based on response processes. Psicothema 2014;26:136-44.

52. Rios J, Wells C. Validity evidence based on internal structure. Psicothema 2014;26:108-16.

53. Popham WJ. Consequential validity: Right Concern-Wrong Concept. Educational Measurement: Issues and Practice 1997;16:9-13.

54. Ma IW, Teteris E, Roberts JM, Bacchus M. Who is teaching and supervising our junior residents' central venous catheterizations? BMC Med Educ 2011;11:16.

55. Velmahos GC, Toutouzas KG, Sillin LF, et al. Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory. Am J Surg 2004;187:114-9.

56. Smith CC, Huang GC, Newman LR, et al. Simulation training and its effect on long-term resident performance in central venous catheterization. Simulation in Healthcare 2010;5:146-51 10.1097/SIH.0b013e3181dd9672.

57. Ahya SN, Barsuk JH, Cohen ER, Tuazon J, McGaghie WC, Wayne DB. Clinical performance and skill retention after simulation-based education for nephrology fellows. Seminars in Dialysis 2012;25:470-3.

58. Madenci AL, Solis CV, de Moya MA. Central venous access by trainees: a systematic review and meta-analysis of the use of simulation to improve success rate on patients. Simulation in Healthcare 2014;9:7-14.

59. Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to their Development and Use. 4th ed. New York: Oxford University Press; 2008.

60. McKinley RK, Strand J, Ward L, Gray T, Alun-Jones T, Miller H. Checklists for assessment and certification of clinical procedural skills omit essential competencies: a systematic review. Medical Education 2008;42:338-49.

61. Evans LV, Dodge KL. Simulation and patient safety: evaluative checklists for central venous catheter insertion. Qual Saf Health Care 2010;19 Suppl 3:i42-6.

62. Bould MD, Crabtree NA, Naik VN. Assessment of procedural skills in anaesthesia. British Journal of Anaesthesia 2009;103:472-83.

63. Reichenheim ME. Confidence intervals for the kappa statistic. Stata Journal 2004;4:421-8.

64. Britt RC, Novosel TJ, Britt LD, Sullivan M. The impact of central line simulation before the ICU experience. Am J Surg 2009;197:533-6.

65. Huang GC, Newman LR, Schwartzstein RM, et al. Procedural competence in internal medicine residents: validity of a central venous catheter insertion assessment instrument. Acad Med 2009;84:1127-34.

66. Lee AC, Thompson C, Frank J, et al. Effectiveness of a novel training program for emergency medicine residents in ultrasound-guided insertion of central venous catheters. Cjem 2009;11:343-8.

67. Millington SJ, Wong RY, Kassen BO, Roberts JM, Ma IW. Improving internal medicine residents' performance, knowledge, and confidence in central venous catheterization using simulators. Journal of hospital medicine : an official publication of the Society of Hospital Medicine 2009;4:410-6.

68. Murphy MA, Neequaye S, Kreckler S, Hands LJ. Should we train the trainers? Results of a randomized trial. Journal of the American College of Surgeons 2008;207:185-90.

69. Ramakrishna G, Higano ST, McDonald FS, Schultz HJ. A curricular initiative for internal medicine residents to enhance proficiency in internal jugular central venous line placement. Mayo Clinic Proceedings 2005;80:212-8.

70. Rosen BT, Uddin PQ, Harrington AR, Ault BW, Ault MJ. Does personalized vascular access training on a nonhuman tissue model allow for learning and retention of central line placement skills? Phase II of the procedural patient safety initiative (PPSI-II). Journal of hospital medicine : an official publication of the Society of Hospital Medicine 2009;4:423-9.

71. Barsuk JH, Ahya SN, Cohen ER, McGaghie WC, Wayne DB. Mastery learning of temporary hemodialysis catheter insertion by nephrology fellows using simulation technology and deliberate practice. Am J Kidney Dis 2009;54:70-6.

72. Barsuk JH, McGaghie WC, Cohen ER, Balachandran JS, Wayne DB. Use of simulation-based mastery learning to improve the quality of central venous catheter placement in a medical intensive care unit. Journal of hospital medicine : an official publication of the Society of Hospital Medicine 2009;4:397-403.

73. Evans LV, Morse JL, Hamann CJ, Osborne M, Lin Z, D'Onofrio G. The development of an independent rater system to assess residents' competence in invasive procedures. Academic Medicine 2009;84:1135-43 10.097/ACM.0b013e3181acec7c.

74. Wall RJ, Ely EW, Elasy TA, et al. Using real time process measurements to reduce catheter related bloodstream infections in the intensive care unit. Quality and Safety in Health Care 2005;14:295-302.

75. Dong Y, Suri HS, Cook DA, et al. Simulation-based objective assessment discerns clinical proficiency in central line placement: a construct validation. Chest 2010;137:1050-6.

76. Blaivas M, Adhikari S. An unseen danger: Frequency of posterior vessel wall penetration by needles during attempts to place internal jugular vein central catheters using ultrasound guidance *. Critical Care Medicine 2009;37:2345-9 10.1097/CCM.0b013e3181a067d4.

77. Carvalho P. Early introduction to central line placement: a curriculum for medical students. Med Educ 2007;41:1098-9.

78. Stone MB, Moon C, Sutijono D, Blaivas M. Needle tip visualization during ultrasound-guided vascular access: short-axis vs long-axis approach. The American Journal of Emergency Medicine 2010;28:343-7.

79. Kilbourne MJ, Bochicchio GV, Scalea T, Xiao Y. Avoiding common technical errors in subclavian central venous catheter placement. Journal of the American College of Surgeons 2009;208:104-9.

80. Lobo RD, Levin AS, Brasileiro Gomes LM, et al. Impact of an educational program and policy changes on decreasing catheter-associated bloodstream infections in a medical intensive care unit in Brazil. American Journal of Infection Control 2005;33:83-7.

81. Coopersmith CM, Rebmann TL, Zack JE, et al. Effect of an education program on decreasing catheter-related bloodstream infections in the surgical intensive care unit. Crit Care Med 2002;30:59-64.

82. Xiao Y, Seagull FJ, Bochicchio GV, et al. Video-based training increases sterile-technique compliance during central venous catheter insertion. Crit Care Med 2007;35:1302-6.

83. Yilmaz G, Caylan R, Aydin K, Topbas M, Koksal I. Effect of education on the rate of and the understanding of risk factors for intravascular catheter-related infections. Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America 2007;28:689-94.

84. McKee C, Berkowitz I, Cosgrove SE, et al. Reduction of catheter-associated bloodstream infections in pediatric patients: Experimentation and reality. Pediatric Critical Care Medicine 2008;9:40-6.

85. Papadimos T, Hensely S, Duggan J, et al. Intensivist supervision of resident-placed central venous catheters decreases the incidence of catheter-related blood stream infections. Patient Safety in Surgery 2008;2:11.

86. Costello JM, Morrow DF, Graham DA, Potter-Bynoe G, Sandora TJ, Laussen PC. Systematic intervention to reduce central line–associated bloodstream infection rates in a pediatric cardiac intensive care unit. Pediatrics 2008;121:915-23.

87. Guidelines for developing evaluation checklists: the checklists development checklists (CDC). Western Michigan University, 2000. (Accessed September 27, 2010, at http://www.wmich.edu/evalctr/checklists/checklist-creation/.)

88. Pugh D, Hamstra SJ, Wood TJ, et al. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. Adv Health Sci Educ Theory Pract 2014.

89. Cunnington JPW, Neville AJ, Norman GR. The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. Advances in Health Sciences Education 1996;1:227-33.

90. Chen FM, Bauchner H, Burstin H. A call for outcomes research in medical education. Academic Medicine 2004;79:955-60.

91. Dath D, Regehr G, Birch D, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. Surgical Endoscopy 2004;18:1800-4.

92. Beard JD, Jolly BC, Newble DI, Thomas WEG, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. British Journal of Surgery 2005;92:778-82.

93. Beckmann CRB, Lipscomb GH, Ling FW, Beckmann CA, Johnson H, Barton L. Computer-Assisted Video Evaluation of Surgical Skills. Obstetrics & Gynecology 1995;85:1039-41.

94. Vogt VY, Givens VM, Keathley CA, Lipscomb GH, Summitt RL. Is a resident's score on a videotaped objective structured assessment of technical skills affected by revealing the resident's identity? American journal of obstetrics and gynecology 2003;189:688-91.

95. Stroud L, Herold J, Tomlinson G, Cavalcanti RB. Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. Academic Medicine 2011;86:S8-S11 0.1097/ACM.0b013e31822a729d.

96. Fuchs D. Examiner familiarity effects on test performance: implications for training and practice. Topics in Early Childhood Special Education 1987;7:90-104.

97. Fuchs D, Fuchs LS. Test procedure bias: A meta-analysis of examiner familiarity effects. Review of Educational Research 1986;56:243-62.

98. Vogt VY, Givens VM, Keathley CA, Lipscomb GH, Summitt Jr RL. Is a resident's score on a videotaped objective structured assessment of technical skills affected by revealing the resident's identity? American journal of obstetrics and gynecology 2003;189:688-91.

99. Scott DJ, Rege RV, Bergen PC, et al. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech A 2000;10:183-90.

100. Vassiliou MC, Feldman LS, Fraser SA, et al. Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. Surg Innov 2007;14:211-6.

101. Kneebone R, Nestel D, Yadollahi F, et al. Assessing procedural skills in context: exploring the feasibility of an Integrated Procedural Performance Instrument (IPPI). Medical Education 2006;40:1105-14.

102. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward Feasible, Valid, and Reliable Video-Based Assessments of Technical Surgical Skills in the Operating Room. Annals of Surgery 2008;247:372-9 10.1097/SLA.0b013e318160b371.

103. House JB, Dooley-Hash S, Kowalenko T, et al. Prospective comparison of live evaluation and video review in the evaluation of operator performance in a pediatric emergency airway simulation. Journal of Graduate Medical Education 2012;4:312-6.

104. Laeeq K, Infusino S, Lin SY, et al. Video-based assessment of operative competency in endoscopic sinus surgery. American journal of rhinology & allergy 2010;24:234-7.

105. Chang L, Hogle NJ, Moore BB, et al. Reliable Assessment of Laparoscopic Performance in the Operating Room Using Videotape Analysis. Surgical Innovation 2007;14:122-6.

106. Driscoll PJ, Paisley AM, Paterson-Brown S. Video assessment of basic surgical trainees' operative skills. The American Journal of Surgery 2008;196:265-72.

107. Williams JB, McDonough MA, Hilliard MW, Williams AL, Cuniowski PC, Gonzalez MG. Intermethod Reliability of Real-time Versus Delayed Videotaped Evaluation of a High-

fidelity Medical Simulation Septic Shock Scenario. Academic Emergency Medicine 2009;16:887-93.

108. Direct Observation of Procedural Skills (DOPS). The Foundation Programme. The UK Foundation Programme Office (Accessed September 24, 2013, at http://www.foundationprogramme.nhs.uk/pages/home/curriculum-and-assessment/curriculum2012.)

109. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. Am J Surg 1997;173:226-30.

110. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.

111. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

112. Oren C, Kennet-Cohen T, Turvall E, Allalouf A. Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. Psicothema 2014;26:117-26.

113. Ringsted C, Østergaard D, Ravn L, Pedersen J, Berlac P, Van der Vleuten C. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. Medical Teacher 2003;25:654-8.

114. Cohen DS, Colliver JA, Robbs RS, Swartz MH. A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardized-patient examination. Advances in Health Sciences Education 1996;1:209-13.

115. Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. Academic Medicine 1996;71:S87-9.

116. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Medical Education 2003;37:1012-6.

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

118. Norman GR, Van Der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Medical Education 1991;25:119-26.

119. Van Der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Medical Education 1991;25:110-8.

120. Cattell RB. Scree test for number of factors. Multivariate Behav Res 1966;1:245-76.

121. Spearman C. The proof and measurement of association between two things. The American Journal of Psychology 1904;15:72-101.

122. Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. New York, NY: John Wiley & Sons, Inc.; 2000.

123. Domino KB, Bowdle TA, Posner KL, Spitellie PH, Lee LA, Cheney FW. Injuries and Liability Related to Central Vascular Catheters: A Closed Claims Analysis. Anesthesiology 2004;100:1411-8.

124. Mansfield PF, Hohn DC, Fornage BD, Gregurich MA, Ota DM. Complications and failures of subclavian-vein catheterization. New England Journal of Medicine 1994;331:1735-8.

125. Yudkowsky R, Tumuluru S, Casey P, Herlich N, Ledonne C. A patient safety approach to setting pass/fail standards for basic procedural skills checklists. Simulation in Healthcare 2014;9:277-82 10.1097/SIH.0000000000000044.

126. Yudkowsky R, Park YS, Riddle J, Palladino C, Bordage G. Clinically discriminating checklists versus thoroughness checklists: improving the validity of performance test scores. Acad Med 2014;89:1057-62.

127. Daniels VJ, Bordage G, Gierl MJ, Yudkowsky R. Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an Internal Medicine residency OSCE. Adv Health Sci Educ Theory Pract 2014;19:497-506.

128. De Champlain A, Margolis M, Macmillan M, Klass D. Predicting mastery level on a large-scale standardized patient test: a comparison of case and instrument score-based models using discriminant function analysis. Advances in Health Sciences Education 2001;6:151-8.

129. Sandilands D, Gotzmann A, Roy M, Zumbo BD, De Champlain A. Weighting checklist items and station components on a large-scale OSCE: Is it worth the effort? Medical Teacher 2014;36:585-90.

130. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA 1993;269:1655-60.

131. Verhulst SJ, Colliver JA, Paiva R, Williams RG. A factor analysis study of performance of first-year residents. Academic Medicine 1986;61:132-4.

132. Ponton-Carss A, Hutchison C, Violato C. Assessment of communication, professionalism, and surgical skills in an objective structured performance-related examination (OSPRE): a psychometric study. American Journal of Surgery 2011;202:433-40.

133. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. In. US: American Psychological Association; 1993:363-406.

134. Royal College of Physicians and Surgeons of Canada. Objectives of Training in the Specialty of Internal Medicine. 2011. (Accessed August 30, 2013, at http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000910.pdf.)

135. Wigton RS, Alguire P. The declining number and variety of procedures done by general internists: a resurvey of members of the American College of Physicians. Ann Intern Med 2007;146:355-60.

136. Royal College of Physicians and Surgeons of Canada. Competency by Design: Reshaping Canadian Medical Education. In.

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

138. Wayne DB, Barsuk JH, O'Leary KJ, Fudala MJ, McGaghie WC. Mastery learning of thoracentesis skills by internal medicine residents using simulation technology and deliberate practice. Journal of Hospital Medicine 2008;3:48-54.

139. Dijkstra J, Van der Vleuten C, Schuwirth L. A new framework for designing programmes of assessment. Advances in Health Sciences Education 2010;15:379-93.

140. Lew SR, Page GG, Schuwirth LWT, et al. Procedures for establishing defensible programmes for assessing practice performance. Med Educ 2002;36:936-41.

141. Norcini JJ, McKinley DW. Assessment methods in medical education. Teaching and Teacher Education 2007;23:239-50.

142. Van Der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Medical Education 2005;39:309-17.

143. Dweck CS. Motivational processes affecting learning. American Psychologist 1986;41:1040.

144. Gal I, Ginsburg L. The role of beliefs and attitudes in learning statistics: Towards an assessment framework. Journal of Statistics Education 1994;2:1-15.

145. Feller-Kopman DMD. Ultrasound-guided central venous catheter placement: The new standard of care? . Crit Care Med 2005;33:1875-7.

146. Walzak A, Bacchus M, Schaefer J, et al. Diagnosing technical competence in six bedside procedures using checklists and global rating scales: is it time to abandon checklists? . Acad Med 2014;In Press.

APPENDIX A: STANDARDIZED DATA COLLECTION FORM

Reviewer Initials: _____ Date of data abstraction_____ Study ID #_____AuthorLastName_____

Population studied (check all that apply):
☐ Medical student ☐ PGY-1 ☐ PGY-2 ☐ PGY-3 ☐ Fellow ☐ Attending/consulting physician
☐ Nurses ☐ Physician Assistants ☐ Other_____ ☐ Not described

Specialty of Learners (check all that apply): ☐ Critical Care ☐ Internal Medicine ☐ Surgery
☐ Emergency ☐ Not mentioned ☐ Other _____

Tool used on...(check all that apply):
☐ Simulator ☐ patient ☐ Other
Ultrasound used ☐ Y ☐ N
Sites tested: ☐ IJ ☐ SC ☐ Fem ☐ Not mentioned ☐ Right ☐ Left

Type of simulation used (check all that apply):
☐ partial task trainers ☐ standardized patients ☐ full body task trainers
☐ high fidelity mannequins ☐ virtual reality ☐ computer software

Total number of subjects tested :

Checklist used? ☐ Y ☐ N
If yes, how were items generated?

Number of categories for checklist: ☐ Binary ☐ ____

Evidence of reliability:

☐ Inter-rater _____ ☐ Intra-rater_____ ☐ test-retest _____

☐ Internal/Cronbach's alpha_____

Evidence of Validity: Type?

Global rating Scale Used? ☐ Y ☐ N
If yes, how were items generated?

Number of Likert scale: _____ From _____ to _____
Behaviourally anchored? ☐ Y ☐ N

Evidence of reliability:

☐ Inter-rater _____ ☐ Intra-rater_____ ☐ test-retest _____

☐ Internal/Cronbach's alpha_____

Evidence of Validity: Type?

Version Sep 20_2011

Reviewer Initials: _____ Date of data abstraction_____ Study ID #_____AuthorLastName_____

No.	Item	Procedural Competence	Preparation	Safety	Communication and working with patient	Infection control	Post-procedural care	Team working
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								

Version Sep 20_2011

APPENDIX B: SIMULATION TRAINING PRE- AND POST-COURSE SURVEY

v. Sep 4, 2008. UBC INTERNAL MEDICINE 2008-2009

CENTRAL LINE INSERTION SIMULATION COURSE

Pre-course Evaluation:

1. Please indicate your gender.
   - ☐ M
   - ☐ F

2. Please indicate how many months of ICU you have completed in total as of today
   - ☐ 0
   - ☐ 1
   - ☐ 2
   - ☐ 3 or more

3. Have you ever been taught formally how to place a central line?
   ☐ Yes ☐ No

   If Yes, who taught you? (please check all that apply)

   ☐ An attending ☐ A resident more senior than yourself
   ☐ A resident same level of training as yourself ☐ someone more junior than yourself

   If Yes, was it taught on a

   ☐ Patient ☐ Simulator ☐ Other _____

4. Please disclose your past experience with central line insertion by reporting the number of central lines insertions you have watched or performed at each site. Include experience as a medical student and a resident. Of the ones performed, how many were supervised?

	Femoral Line	IJ Line	SC Line
Number Watched			
Number Performed			
Of the ones performed…how many were			
Supervised by an attending			
Supervised by a resident more senior than yourself			
Supervised by a resident more junior than yourself			
Unsupervised			
(total must add up to no. Performed)			

5. On the following scale, please rate your level of confidence with central line insertion in general
   - ☐ None
   - ☐ Very little
   - ☐ Some
   - ☐ Moderate
   - ☐ Good
   - ☐ Complete

Post-course Evaluation:

1. On the following scale, please rate your level of confidence with central line insertion
   - ☐ None
   - ☐ Very little
   - ☐ Some
   - ☐ Moderate
   - ☐ Good
   - ☐ Complete

2. Overall, how satisfied were you with the central line insertion course?
   - ☐ Very low
   - ☐ Low
   - ☐ Moderate
   - ☐ High
   - ☐ Very high

3. Would you recommend that this course continued to be offered?
   - ☐ No
   - ☐ Yes

4. Please respond to the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
This course improved my technical ability to insert central lines	☐	☐	☐	☐	☐
This course decreased my anxiety level related to line placement	☐	☐	☐	☐	☐
This course improved my ability to deliver patient care.	☐	☐	☐	☐	☐

5. Please rate the following

	Very Poor	Poor	Adequate	Good	Very Good
Actual teaching session	☐	☐	☐	☐	☐
Time for supervised practice	☐	☐	☐	☐	☐
Realism of dummies	☐	☐	☐	☐	☐
CESEI facility overall	☐	☐	☐	☐	☐

Additional Comments:

APPENDIX C: GLOBAL RATING SCALE

Please rate the following areas:	Not competent to perform independently	Borderline competence to perform independently	Competent to perform independently	Above average competence to perform independently			NA
Appropriate preparation of instruments pre-procedure	☐	☐	☐	☐	☐	☐	☐
Appropriate analgesia	☐	☐	☐	☐	☐	☐	☐
Specific components of technical ability: Time and Motion	☐ Many unnecessary moves	☐	☐ Efficient but some unnecessary moves	☐	☐ Economy of movement & maximum efficiency		NA ☐
Instrument handling	☐ Repeatedly makes tentative and awkward moves	☐	☐ Competent, occasionally appeared stiff or awkward	☐	☐ Fluid movements, no awkwardness		NA ☐
Flow of procedure and forward planning	☐ Frequently stopped or needed to discuss next move	☐	☐ Demonstrated ability for forward planning	☐	☐ Obviously planned course with effortless flow		NA ☐
Knowledge of instruments	☐ Frequently used an inappropriate instrument	☐	☐ Used appropriate instrument for the task	☐	☐ Obviously familiar with required instruments		NA ☐
Aseptic technique	☐	☐	☐	☐	☐	☐	☐
Seeks help where appropriate	☐	☐	☐	☐	☐	☐	☐
OVERALL ABILITY TO PERFORM PROCEDURE	☐	☐	☐	☐	☐	☐	☐

APPENDIX D: TEN-ITEM CHECKLIST

Date: _____

Evaluator: _____

Resident: _____

STEP	YES	NO
Justified site selection	NA	NA
Justified catheter selection	NA	NA
1. Prepared site appropriately		
Requested Trendelenberg position	NA	NA
2. Identified proper landmarks		
3. Inserted needle at correct angle		
4. Aspirated while inserting needle		
5. Inserted guidewire appropriately		
6. Withdrew needle and incised skin		
7. Inserted catheter and withdrew wire		
8. Aspirated blood and flushed ports		
9. Occluded ports		
10. Secured catheter		
Sealed site in sterile fashion	NA	NA

*Reproduced with permission from Springer, *from* Ma I et al. Adv Health Sci Educ Theory Pract 2012.

SKILL	RESULT
Number of attempts to locate the vein	

APPENDIX E: TWENTY-ONE ITEM CHECKLIST

Date: _____

Evaluator: _____

Resident: _____

STEP	**YES**	**NO**
Flush the ports on the catheter with sterile saline		
Clamp each port (*ok to keep brown port open)*		
Remove brown port from end of catheter to accommodate wire		
Area is cleaned with chlorhexadine		
Resident gets in sterile gown, gloves, hat and mask		
Area is draped in usual sterile fashion (*must be full body drape)*		
The vein is localized using anatomical landmarks		
The skin is anesthetized with 1% lidocaine in a small wheal		
The deeper structures are anesthetized		
Using the large needle or catheter-syringe complex, cannulate the vein while aspirating		
Remove the syringe from the needle or advance the catheter into the vein removing both the syringe and the needle		
Advance the guidewire into the vein no more than approximately 12-15cm		
Knick the skin with the scalpel to advance the dilator		
Advance the dilator over the guidewire and dilate the vein		
Advance the triple lumen over the guidewire		
Never let go of the guidewire		
Once the catheter is inserted remove the guidewire in its entirety		
Advance the catheter to approx 14-16 cm on the right side		
Ensure there is blood flow/flush each port		
Secure the catheter in place (*suture or staple)*		
Maintain sterile technique		

*Reproduced with permission from Springer, *from* Ma I et al. Adv Health Sci Educ Theory Pract 2012.