

Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation

Irene W. Y. Ma · Nadia Zalunardo · George Pachev · Tanya Beran ·
Melanie Brown · Rose Hatala · Kevin McLaughlin

Received: 18 February 2011 / Accepted: 21 August 2011 / Published online: 30 August 2011
© Springer Science+Business Media B.V. 2011

Abstract The use of checklists is recommended for the assessment of competency in central venous catheterization (CVC) insertion. To explore the use of a global rating scale in the assessment of CVC skills, this study seeks to compare its use with two checklists, within the context of a formative examination using simulation. Video-recorded performances of CVC insertion by 34 first-year medical residents were reviewed by two independent, trained evaluators. Each evaluator used three assessment tools: a ten-item checklist, a 21-item checklist, and a nine-item global rating scale. Exploratory principal component analysis of the global rating scale revealed two factors, accounting for 84.1% of the variance: technical ability and safety. The two checklist scores correlated positively with the weighted factor score on technical ability (0.49 [95% CI 0.17–0.71] for the 10-item checklist; 0.43 [95% CI 0.10–0.67] for the 21-item checklist) and negatively with the weighted factor score on safety (−0.17 [95% CI −0.48–0.18] for the 10-item checklist; −0.13 [95% CI −0.45–0.22] for the 21-item checklist). A checklist score of <80% was strong indication of incompetence. However, a high checklist score did not preclude incompetence. Ratings using the global rating scale identified an additional 11 candidates (32%) who were deemed incompetent despite scoring >80% on both checklists. All these candidates committed serious errors. In conclusion, the practice of universal adoption of checklists as the preferred method of assessment of procedural skills should be questioned. The inclusion of global rating scales should be considered.

I. W. Y. Ma (✉) · K. McLaughlin
Department of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary,
AB T2N 4N1, Canada
e-mail: ima@ucalgary.ca

N. Zalunardo · M. Brown · R. Hatala
Department of Medicine, University of British Columbia, Vancouver, BC, Canada

G. Pachev
Educational Assessment Unit, Faculty of Medicine, University of British Columbia,
Vancouver, BC, Canada

T. Beran
Department of Community Health Sciences and Medical Education Research Unit,
University of Calgary, Calgary, AB, Canada

Keywords Simulation · Catheterization, central venous · Clinical competency · Checklist

Introduction

Central venous catheterization (CVC) is a commonly performed bedside medical procedure. Competency in this procedure is an explicit objective for a number of postgraduate training programs, including emergency medicine, internal medicine, critical care medicine, and general surgery (ACGME 2007; RCPSC 2003, 2005, 2008, 2010; Joint Royal Colleges of Physicians Training Board 2009). The Accreditation Council for Graduate Medical Education (ACGME) recommends the use of simulation and checklists as the “most desirable” evaluation methods for the assessment of competency in procedural skills (ACGME 2000). The use of a global rating scale, on the other hand, is only listed as a “potentially applicable method”. Perhaps related in part to these recommendations, procedural checklists for CVC have become commonly used (Barsuk et al. 2009a; Dong et al. 2010; Evans and Dodge 2010; Velmahos et al. 2004). Indeed, in our systematic review of twenty studies examining the use of simulation-based education for CVC (Ma et al. 2011), only two studies used global rating scales for the evaluation of procedural performances (Lee et al. 2009; Millington et al. 2009).

Despite the frequency with which checklists are used to evaluate CVC skills, the rationale for the recommendation of their use is unclear. A common misconception about checklists is that they are more objective and therefore result in more reliable ratings than global rating scales (Cohen et al. 1996). However, this misconception has been previously challenged (Norman et al. 1991; Van Der Vleuten et al. 1991). The use of checklists has not been shown to consistently result in an improvement in reliability (Cohen et al. 1996; Van Der Vleuten et al. 1991). Furthermore, ratings by experts using global rating scales can outperform checklists in terms of their reliability and validity measures (Hodges and McIlroy 2003; Regehr et al. 1998). Moreover, in objective structured clinical examinations (OSCEs) for clinical skills, unlike global rating scales, checklists have been shown to have low sensitivity to increasing levels of expertise (Hodges et al. 1999; Hodges and McIlroy 2003). It has been postulated that the use of checklists runs the risk of trivializing steps by rewarding thoroughness rather than clinical competence (Cunnington et al. 1996; Norman et al. 1991). Therefore, rather than automatically adopting objectified methods of assessment such as checklists (Van Der Vleuten et al. 1991), the choice of assessment tools should be made based on best available evidence (Norman 2005).

The use of subjective expert judgments in defining competency is not unprecedented. For example, pass/fail scores of the National high stakes OSCE examination for the Licentiate of the Medical Council of Canada (LMCC) are based on experts’ overall judgment on global performance (Dauphinee et al. 1997). Borderline checklist scores are then calculated based on performances of candidates deemed borderline on their overall global performance. Expert judgment using global rating scales has also been previously used in the assessment of surgical skills (Reznick et al. 1997).

To explore the use of a global rating scale in the assessment of bedside CVC skills, this study seeks to compare its use with two checklists, within the context of a formative examination using simulation. To do so, we first explored the dimensions captured by our constructed global rating scale. We then evaluated the correlations of scores obtained among the different tools. Lastly, we evaluated the diagnostic performance of checklist

scores in identifying competence, based on expert physician global judgment of candidates' performances.

Method

Participants

During the academic year of 2008–2009, all first year internal medicine residents at the University of British Columbia, who provided written informed consent, were included in the study. The study was approved by our university ethics review board.

Participants were enrolled in a 2-h simulator training session on CVC. Details of this curriculum involving a different cohort of participants have been previously described (Millington et al. 2009). At the end of the simulator training session, the participants underwent a formative examination using simulators. The examination consisted of a performance of an internal jugular CVC on a simulator (Laerdal IV Torso; Laerdal Medical Corp, Wappingers Falls, New York) using a standard kit provided. Participants were instructed to perform a CVC as they would in real-life, without externally imposed time limits on the procedure. Feedback about their examination 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-taped in a blinded fashion, with no personal identifying information recorded.

Evaluation tools

Three tools were used for this study. The global rating scale is an eight-item scale, with an additional ninth summary item on “overall ability to perform procedure” (“Appendix 1”). The eight items were adapted from two validated global rating scales: Direct Observation of Procedural Skills (DOPS; The Foundation Programme 2009) and the Objective Structured Assessment of Technical Skills (OSATS; Reznick et al. 1997). Items on the original scales not applicable to our simulator examination were removed. After piloting our assessment tool, this rating scale was 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”. We dichotomized the scores for this item such that a score of three or more was considered “competent to perform the procedure”, while a score of two or less was considered “not competent to perform the procedure”.

Two checklists were used for this study. The first checklist (“Appendix 2”) consists of ten items, adapted from a previously published checklist (Velmahos et al. 2004).

The second checklist (“Appendix 3”) consists of 21 items, adapted from a previously validated twenty-seven item checklist (Barsuk et al. 2009b). This checklist was published after the completion of our initial simulation assessments and was included in our study after it was made available. Items on both original checklists deemed not applicable to our simulator examination were removed. The overall checklist score for each checklist was calculated as the number of completed items divided by the total number of items, presented as a percentage.

Rather than assuming validity of our modified assessment tools, content validity of the final global rating scale and checklists 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.

Video performance evaluation

All video-recorded performances were evaluated by two independent trained evaluators who are faculty members with experience in simulator teaching. Evaluators were trained for 3 h on the use of each assessment tool, by review of four videos recorded specifically for the purposes of training. After training, the intraclass correlation coefficients of the evaluators were >0.80 . For the evaluation of video-recorded performances, for each evaluator, 50% of the videos were rated first using the ten-item checklist while the remaining 50% of the videos were rated first using the global rating scale.

To assess the extent to which one tool may have systematically influenced the rating of a subsequent tool, the same raters re-analyzed each video approximately 2 years after the completion of the study. During the re-analysis, all assessment tools that were initially rated after 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 intraclass correlation coefficients for the global rating between ratings with and without the influence of another tool was 0.92. Overall Kappa score for the ten-item checklist was 0.92, with a summary checklist score reliability of 0.97. Evaluation using the 21-item checklist was done independently by two raters 1 year after the initial evaluation and therefore was not subject to the influence of rating by another tool.

Statistical analysis

To explore the dimensions assessed by the global rating scale, the following analyses were performed: after confirming the appropriateness of performing factor 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 components analysis was performed on the eight items in the global rating scale, using a VARIMAX rotation. A Scree plot was inspected (Cattell 1966). Factors with eigenvalues greater than 1 were retained. Item loadings ≥ 0.40 are reported. Inter-rater reliability was evaluated using intraclass correlation coefficient (ICC), Pearson's correlation coefficient, and Cohen's Kappa where appropriate. Both correlation coefficients and disattenuated coefficients are reported (Spearman 1904). 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 three 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

Thirty-five participants were invited and 34 (97%) consented to and completed the study protocol (Table 1).

Dimensions assessed by the global rating scale

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 2). The first factor consisted primarily of

Table 1 Participants' demographic characteristics (total $N = 34$)

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

IQR interquartile range

Table 2 Rotated factor loadings for scale items

Global rating scale item	Factor 1—technical ability	Factor 2—safety
Knowledge of instruments	0.93	
Instrument handling	0.90	
Flow of procedure and forward planning	0.90	
Time and motion	0.88	
Appropriate analgesia	0.47	
Seeks help where appropriate		0.87
Aseptic technique		0.59
Appropriate preparation of instruments pre-procedure		

Factor loadings below 0.40 are suppressed

behaviors that can be characterized as relating to “technical ability” ($\alpha = 0.78$). The second factor consisted of behaviors that relate to procedural “safety” ($\alpha = 0.76$). The item “appropriate preparation of instruments pre-procedure” did not load on any factor.

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

Inter-rater reliability of individual items and overall rating of the global rating scale and the two checklists is shown in Table 3.

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

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 0.10–0.67). The correlation between the 21-item checklist score and the weighted factor score on procedural safety was −0.13 (95% confidence interval −0.45–0.22).

Diagnostic performance of checklist scores in identifying competence based on expert global judgment

Based on expert global judgment of competence, twenty-one participants (62%) were rated overall as being competent, while 13 (38%) were rated as not. There were no significant baseline differences between the two groups (Table 1).

The mean overall score on the ten-item checklist for those deemed competent was $95.2 \pm 8.1\%$, which is significantly higher compared to those who were deemed not competent ($81.0 \pm 18.6\%$, $p = 0.002$). Correlation between the overall ten-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 ($92.0 \pm 5.2\%$ vs. $84.1 \pm 14.7\%$ 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.

On ROC analyses, the overall ten-item checklist score had an acceptable discrimination (AUC = 0.79, standard error = 0.078, 95% confidence interval [0.64, 0.94]) (Hosmer and Lemeshow 2000), while the 21-item checklist’s AUC was 0.68, standard error = 0.098, 95% confidence interval [0.48, 0.87]). Table 4 shows the sensitivity and specificity for different cut-off points for the checklist score. For maximum sensitivity (100%), a cut-off point of 80% was chosen as the optimal cut point 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.

Out of the 13 participants who were deemed incompetent on the global rating scale, 11 individuals scored $\geq 80\%$ on the checklists. Reasons for incompetence in these 11 participants included significant breaches in sterility ($n = 5$), loss of wire control ($n = 4$; median duration of time without wire control 35 s, range 17–38 s), multiple attempts ($n = 2$; both cases >6 attempts were made).

Table 3 Inter-rater Reliability (intraclass correlation coefficients or Kappa statistics) for items in the global rating scale and the checklist

	Intraclass correlation coefficient	95% confidence Interval
Global rating scale items ($\alpha = 0.79$) ^a		
Appropriate preparation of instruments pre-procedure	-0.04	-0.38 to 0.30
Appropriate analgesia	0.67	0.42-0.82
Time and motion	0.52	0.23-0.73
Instrument handling	0.63	0.33-0.80
Flow of procedure and forward planning	0.30	-0.03 to 0.57
Knowledge of instruments	0.44	0.12-0.67
Aseptic technique	0.62	0.36-0.79
Seeks help where appropriate	0.63	0.37-0.80
Overall ability to perform procedure	0.63	0.38-0.80
Overall agreement on competence	Kappa = 0.52	0.23-0.81
	Kappa statistic	95% confidence interval
Ten-item checklist ($\alpha = 0.67$)		
Prepared site appropriately	0.77	0.46-1.00
Identified proper landmarks	Perfect agreement	
Inserted needle at correct angle	0.39	0.00-0.78
Aspirated while inserting needle	NA ^b	
Inserted guidewire appropriately	0.65	0.03-1.00
Withdrew needle and incised skin	0.79	0.38-1.00
Inserted catheter and withdrew wire	Perfect agreement	
Aspirated blood and flushed ports	0.64	0.18-1.00
Occluded ports	0.62	0.24-1.00
Secured catheter	0.84	0.54-1.00
Inter-rater reliability of overall ten-item checklist score	$r = 0.88; p < 0.001$	
Twenty-one item checklist ($\alpha = 75$)		
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 (<i>must be full body drape</i>)	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	

Table 3 continued

	Kappa statistic	95% confidence interval
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 12–15 cm	0.53	0.14–0.93
Knock 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	$r = 0.89; p < 0.001$	

^a α refers to Cronbach's alpha, an assessment of internal reliability

^b Not applicable (no variability in rater 2)

Table 4 Sensitivity and specificity of various cutpoints on the overall checklist scores in determining competence

Cut point 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	7.7	7.7
≥ 80	100	100	23.1	23.1
≥ 90	81.0	76.2	53.9	53.9
100	28.6	9.5	15.4	100

Competence is defined by a rating of “competent to perform independently” or better on the summary measure in the Global Rating Scale

Discussion

For the assessment of competence of CVC in a formative examination using simulators, this study evaluated 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. First, with respect to content validity, our results indicate that while dimensions captured by the global rating

scale were technical ability and safety, neither checklist adequately captured errors relating to safety issues. This finding is consistent with the literature on procedural checklists in general. A systematic review on procedural checklists found that 30–50% of checklists did not assess for competencies in the area of ‘infection control’ and ‘safety’ (McKinley et al. 2008). Errors identified by our study 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% (McGee and Gould 2003). Meticulous attention to sterility is an important aspect of the procedure. In a study evaluating malpractice claims for CVC, the most common complication was wire/catheter embolism (Domino et al. 2004). Wire control, therefore, has important safety implications. Lastly, the incidence of mechanical complications increases significantly with three or more attempts at insertion (Mansfield et al. 1994; McGee and Gould 2003). 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 checklist scores with the technical ability factor score on the global rating scale. However, 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 direction, thereby lending support to the impression that these checklists perhaps capture items relating to technical ability more than they do on safety parameters.

Lastly, our results indicate that the use of checklist scores in predicting 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 ($\geq 80\%$) were still deemed incompetent. All of these candidates committed errors that were considered serious in nature.

What are the implications of these results? Consistent with the literature on evaluations for OSCE, our results suggest that checklists should not automatically be assumed to be the preferred method of assessment. While we do not advocate that checklists be abandoned altogether for the assessment of procedural skills, we do however recommend that their use be evaluated prior to their adoption.

Both of the checklists evaluated in our study were constructed carefully; one used a cognitive task analysis approach (Velmahos et al. 2004), while the other used a rigorous checklist development procedure (Barsuk et al. 2009b). Nonetheless, despite careful construction, improvement in content validation may be made to these tools either by including additional items of safety parameters or differentially weighting critical items.

Limitations

Our 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 easily argue for the use of checklist scores as the standard instead.

Our study chose the use of a global rating scale as the standard to maximize the number of trainees identified as potentially benefiting from additional practice. For a formative examination, 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 a global rating scale identified additional incompetent performances that were rated highly by both checklist scores. Furthermore, these additional individuals identified in our study sample did not appear to be a result of a false positive identification in that all these individuals committed serious errors that were considered to pose significant harm to patients.

Thirdly, results of our study conclusions regarding the use of assessment tools are context-specific. For example, our conclusions relate to the two checklists and the global rating scale used in our study, in a formative simulation 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 using ultrasound technique (NHS 2002). 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 including additional items on safety parameters, differentially weighting critical items, or the use a three-point scale for each item on the checklists rather than their current binary form.

Finally, the validity of score from our constructed global rating scale cannot be assumed, despite the fact that it was constructed based on two previously validated tools (The Foundation Programme 2009; Reznick et al. 1997). 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 cut-point for competence was chosen arbitrarily, albeit by consensus, based on concerns that evaluations may be positively skewed (Streiner and Norman 2003). As a result, three categories were available to assist examiners in differentiating among above average performances. The trained evaluators in our study, however, ultimately deemed 38% of the candidates as incompetent. 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 two currently available checklists. Future study should focus on further optimizing the construction of assessment tools and correlating assessment results with clinical outcomes.

Acknowledgments This work was presented in part at the 2009 Royal College of Physicians and Surgeons of Canada International Conference on Residency Education in Victoria, BC, CANADA. This work is funded by the Department of Medicine, University of British Columbia and Department of Medicine, University of Calgary. The funding agencies had no role in the design and conduct of this study; in the collection, management, analysis, and interpretation of the data; or in the preparation, review, and approval of the manuscript. We thank Dr. Mary Brindle for her assistance with the evaluations of the videos.

Appendix 1: 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	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Appropriate analgesia	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Specific components of technical ability:							NA
Time and Motion	<input type="checkbox"/> Many unnecessary moves	<input type="checkbox"/>	<input type="checkbox"/> Efficient but some unnecessary moves	<input type="checkbox"/>	<input type="checkbox"/> Economy of movement & maximum efficiency		<input type="checkbox"/>
Instrument handling	<input type="checkbox"/> Repeatedly makes tentative and awkward moves	<input type="checkbox"/>	<input type="checkbox"/> Competent, occasionally appeared stiff or awkward	<input type="checkbox"/>	<input type="checkbox"/> Fluid movements, no awkwardness		NA <input type="checkbox"/>
Flow of procedure and forward planning	<input type="checkbox"/> Frequently stopped or needed to discuss next move	<input type="checkbox"/>	<input type="checkbox"/> Demonstrated ability for forward planning	<input type="checkbox"/>	<input type="checkbox"/> Obviously planned course with effortless flow		NA <input type="checkbox"/>
Knowledge of instruments	<input type="checkbox"/> Frequently used an inappropriate instrument	<input type="checkbox"/>	<input type="checkbox"/> Used appropriate instrument for the task	<input type="checkbox"/>	<input type="checkbox"/> Obviously familiar with required instruments		NA <input type="checkbox"/>
Aseptic technique	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Seeks help where appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
OVERALL ABILITY TO PERFORM PROCEDURE	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix 2: Ten-item checklist

Date: _____

Evaluator: _____

Resident: _____

Procedural Checklist

STEP	YES	NO
Justified site selection	NA	NA
Justified catheter selection	NA	NA
Prepared site appropriately		
Requested trendelenberg position	NA	NA
Identified proper landmarks		
Inserted needle at correct angle		
Aspirated while inserting needle		
Inserted guidewire appropriately		
Withdrew needle and incised skin		
Inserted catheter and withdrew wire		
Aspirated blood and flushed ports		
Occluded ports		
Secured catheter		
Sealed site in sterile fashion	NA	NA

Appendix 3: Twenty-one item checklist

STEP	YES	NO
Flush the ports on the catheter with sterile saline		
Clamp each port (<i>ok to keep brown port open</i>)		
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 (<i>must be full body drape</i>)		
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 12–15 cm		
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 (<i>suture or staple</i>)		
Maintain sterile technique		

References

- ACGME. (2000). Accreditation Council for Graduate Medical Education Competencies: Suggested Best Methods for Evaluation. <http://www.acgme.org/Outcome/assess/ToolTable.pdf>. Accessed 17 February 2011.
- ACGME. (2007). Accreditation Council for Graduate Medical Education Program Requirements for Graduate Medical Education in Critical Care Medicine. http://www.acgme.net/acWebsite/downloads/RRC_progReq/142pr707_ims.pdf. Accessed 17 February 2011.
- Barsuk, J. H., Ahya, S. N., Cohen, E. R., McGaghie, W. C., & Wayne, D. B. (2009a). Mastery learning of temporary hemodialysis catheter insertion by nephrology fellows using simulation technology and deliberate practice. *American Journal of Kidney Diseases*, *1*(54), 70–76.
- Barsuk, J. H., Cohen, E. R., Feinglass, J., McGaghie, W. C., & Wayne, D. B. (2009b). Use of simulation-based education to reduce catheter-related bloodstream infections. *Archives of Internal Medicine*, *169*(15), 1420–1423.
- Cattell, R. B. (1966). Scree test for number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.
- Cohen, D. S., Colliver, J. A., Robbs, R. S., & Swartz, M. H. (1996). 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*, *1*(3), 209–213.
- Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1996). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education*, *1*(3), 227–233.
- Dauphinee, W. D., Blackmore, D. E., Smee, S., Rothman, A. I., & Reznick, R. (1997). Using judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Advances in Health Sciences Education*, *2*, 201–211.
- Domino, K. B., Bowdle, T. A., Posner, K. L., Spitellic, P. H., Lee, L. A., & Cheney, F. W. (2004). Injuries and liability related to central vascular catheters: A closed claims analysis. *Anesthesiology*, *100*(6), 1411–1418.
- Dong, Y., Suri, H. S., Cook, D. A., Kashani, K. B., Mullon, J. J., Enders, F. T., et al. (2010). Simulation-based objective assessment discerns clinical proficiency in central line placement: A construct validation. *Chest*, *137*(5), 1050–1056.
- Evans, L. V., & Dodge, K. L. (2010). Simulation and patient safety: Evaluative checklists for central venous catheter insertion. *Quality and Safety in Health Care*, *19*(Suppl 3), i42–i46.
- Foundation Programme. (2009). The Foundation Learning Portfolio. <http://www.foundationprogramme.nhs.uk/pages/home/key-documents#curriculum>. Accessed 17 February 2011.
- Hodges, B., & McLroy, J. H. (2003). Analytic global OSCE ratings are sensitive to level of training. *Medical Education*, *37*, 1012–1016.
- Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine*, *74*(10), 1129–1134.
- Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression*. New York, NY: Wiley.
- Joint Royal Colleges of Physicians Training Board. (2009). Specialty training curriculum for general internal medicine. http://www.gmc-uk.org/2009_GIM_curriculum_FINAL__2_.pdf_30562900.pdf. Accessed 17 February 2011.
- Lee, A. C., Thompson, C., Frank, J., Beecker, J., Yeung, M., Woo, M. Y., et al. (2009). Effectiveness of a novel training program for emergency medicine residents in ultrasound-guided insertion of central venous catheters. *Canadian Journal of Emergency Medicine*, *11*(4), 343–348.
- Ma, I. W., Brindle, M. E., Ronksley, P. E., Lorenzetti, D. L., Sauve, R. S., & Ghali, W. A. (2011). Use of simulation-based education to improve outcomes of central venous catheterization: A systematic review and meta-analysis. *Academic Medicine*, *86*(9), 1137–1147.
- Mansfield, P. F., Hohn, D. C., Fornage, B. D., Gregurich, M. A., & Ota, D. M. (1994). Complications and failures of subclavian-vein catheterization. *New England Journal of Medicine*, *331*(26), 1735–1738.
- McGee, D. C., & Gould, M. K. (2003). Preventing complications of central venous catheterization. *New England Journal of Medicine*, *348*(12), 1123–1133.
- McKinley, R. K., Strand, J., Ward, L., Gray, T., Alun-Jones, T., & Miller, H. (2008). Checklists for assessment and certification of clinical procedural skills omit essential competencies: A systematic review. *Medical Education*, *42*(4), 338–349.
- Millington, S. J., Wong, R. Y., Kassen, B. O., Roberts, J. M., & Ma, I. W. (2009). Improving internal medicine residents' performance, knowledge, and confidence in central venous catheterization using simulators. *Journal of Hospital Medicine*, *4*(7), 410–416.

- NHS. (2002). National Institute for clinical excellence. Guidance on the use of ultrasound locating devices for placing central venous catheters. <http://www.nice.org.uk/nicemedia/live/11474/32462/32462.pdf>. Accessed 18 February 2011.
- Norman, G. (2005). Editorial—Checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Advances in Health Sciences Education*, *10*, 1–3.
- Norman, G. R., Van Der Vleuten, C. P. M., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education*, *25*, 119–126.
- RCPSC. (2003). The Royal College of Physicians and Surgeons of Canada Objectives of Training in Internal Medicine. http://www.rcpsc.medical.org/residency/certification/objectives/intmed_e.pdf. Accessed 17 February 2011.
- RCPSC. (2005). The Royal College of Physicians and Surgeons of Canada Objectives of Training in Adult Critical Care Medicine. http://www.rcpsc.medical.org/residency/certification/objectives/criticalcare-adu_e.pdf. Accessed 17 February 2011.
- RCPSC. (2008). The Royal College of Physicians and Surgeons of Canada Objectives of Training in Emergency Medicine. http://www.rcpsc.medical.org/residency/certification/objectives/emergmed_e.pdf. Accessed 17 February 2011.
- RCPSC. (2010). The Royal College of Physicians and Surgeons of Canada Objectives of Training in the Specialty of General Surgery. http://www.rcpsc.medical.org/residency/certification/objectives/gen_surg_e.pdf. Accessed 17 February 2011.
- Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, *73*(9), 993–997.
- Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative “bench station” examination. *The American Journal of Surgery*, *173*(3), 226–230.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.
- Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales. A practical guide to their development and use* (Third Edition ed.). New York: Oxford University Press Inc.
- Van Der Vleuten, C. P. M., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education*, *25*, 110–118.
- Velmahos, G. C., Toutouzas, K. G., Sillin, L. F., Chan, L., Clark, R. E., Theodorou, D., et al. (2004). Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory. *The American Journal of Surgery*, *187*(1), 114–119.