Use of an error-focused checklist to identify incompetence in lumbar puncture performances

Irene W Y Ma,1,2 Debra Pugh,3 Briseida Mema,4 Mary E Brindle,5 Lara Cooke6 & Julie N Stromer2

CONTEXT Checklists are commonly used in the assessment of procedural competence. However, on most checklists, high scores are often unable to rule out incompetence as the commission of a few serious procedural errors typically results in only a minimal reduction in performance score. We hypothesised that checklists constructed based on procedural errors may be better at identifying incompetence.

OBJECTIVES This study sought to compare the efficacy of an error-focused checklist and a conventionally constructed checklist in identifying procedural incompetence.

METHODS We constructed a 15-item error-focused checklist for lumbar puncture (LP) based on input from 13 experts in four Canadian academic centres, using a modified Delphi approach, over three rounds of survey. Ratings of 18 video-recorded performances of LP on simulators using the error-focused tool were compared with ratings obtained using a published conventional 21-item checklist. Competence/incompetence decisions were based on global assessment. Diagnostic accuracy was estimated using the area under the curve (AUC) in receiver operating characteristic analyses.

RESULTS The accuracy of the conventional checklist in identifying incompetence was low (AUC 0.11, 95% confidence interval [CI] 0.00–0.28) in comparison with that of the error-focused checklist (AUC 0.85, 95% CI 0.67–1.00). The internal consistency of the error-focused checklist was lower than that of the conventional checklist (α = 0.35 and α = 0.79, respectively). The inter-rater reliability of both tools was high (conventional checklist: intraclass correlation coefficient [ICC] 0.99, 95% CI 0.98–1.00; error-focused checklist: ICC 0.92, 95% CI 0.68–0.98).

CONCLUSIONS Despite higher internal consistency and inter-rater reliability, the conventional checklist was less accurate at identifying procedural incompetence. For assessments in which it is important to identify procedural incompetence, we recommend the use of an error-focused checklist.

Medical Education 2015: 49: 1004–1015
doi: 10.1111/medu.12809
Discuss ideas arising from the article at www.mededuc.com discuss.

1Department of Medicine, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada
2W21C, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada
3Department of Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada
4Department of Critical Care Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
5Department of Surgery, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada
6Department of Clinical Neurosciences, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada

Correspondence: Irene W Y Ma, Department of Medicine, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, Alberta T2N 4N1, Canada. Tel: 00 1 403 210 7369; E-mail: ima@ucalgary.ca

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
INTRODUCTION

Medical trainees from a number of postgraduate training programmes are expected to demonstrate competence in the performance of bedside procedures that are critical and essential to patient care.\textsuperscript{1–6} To assess for skills competence, educators commonly turn to two types of assessment tool: the checklist and the global rating scale. The checklist rates observable behaviours typically in a stepwise fashion, whereas the global rating scale tends to rate performances based on an overall or global impression. Both have been shown to have good inter-rater reliability.\textsuperscript{7}

In the case of central venous catheterisation, the number of checklists in use in the literature far outnumbers that of global rating scales.\textsuperscript{8,9} The reason why checklists are used more commonly than global ratings for some procedural skills is not entirely clear. As procedural steps are typically executed in a predictable stepwise fashion, it is conceivable that checklists are felt to be better suited to the assessment task.\textsuperscript{10} Further, items on the checklist may assist in providing feedback to trainees. However, existing data suggest that global rating scales may demonstrate inter-station reliability and validity superior to those of checklists.\textsuperscript{7,11–15} Further, by ‘rewarding thoroughness’, checklists may run the risk of trivialising the task at hand,\textsuperscript{13–15} which may account for their lower ability to distinguish between novices and experts, as demonstrated by one study on non-technical skills.\textsuperscript{16}

In our previous studies of the assessment of competence in procedural skills, we found that the use of checklists resulted in high sensitivity, but low specificity in the identification of competence, a finding that held true across a number of procedures.\textsuperscript{17,18} Particularly worrisome was the finding that a high checklist score did not preclude procedural incompetence as one serious procedural error would result in only a minimal loss in the checklist score if the remaining steps were executed perfectly. This finding suggests that there may be room for improvement in the way that checklists are constructed.

More recently, studies on the inclusion of clinically discriminating items in checklists demonstrate that these generate psychometric data superior to those of conventionally constructed checklists.\textsuperscript{19–21} Thus, the type of item used in a checklist may be an important facet to explore for improving checklist content. Based on findings from our previous studies that procedural errors appeared to account for the poor specificities demonstrated by conventionally constructed checklists,\textsuperscript{17,18} the present study sought to compare the use of a conventionally constructed checklist with that of an error-focused checklist in lumbar puncture (LP). We chose LP because it is a commonly performed procedure that is a requirement for a number of training programmes.\textsuperscript{2,4,6} We hypothesised that a checklist that considers serious procedural errors would outperform a conventionally constructed checklist in its determination of procedural incompetence.

METHODS

Development of the conventional checklist

The development of our 21-item conventionally constructed checklist (Appendix S1) has been described previously.\textsuperscript{18} In short, one neurologist, one emergency medicine specialist, one general internist, two haematologists and one anaesthesiologist from two tertiary Canadian academic centres (University of Calgary and University of British Columbia) participated in an expert panel by completing two rounds of surveys online between December 2010 and October 2011. Consensus, defined as agreement of at least 80%, on the 21 checklist items was reached in a modified Delphi approach.\textsuperscript{22} This checklist was analysed in a binary approach: a score of 1 was assigned for behaviours observed and correctly performed (rated as ‘Yes’), and a score of 0 was assigned for behaviours not observed (rated as ‘No’) or incorrectly performed (rated as ‘Yes, but’).

Development of a conventional global rating tool

The conventional global rating tool has also been described previously.\textsuperscript{17,18} This tool has one item on each of the following domains: pre-procedure preparation; analgesia; time and motion; instrument handling; procedural flow; knowledge of instruments; aseptic technique, and seeking help (Appendix S1). These domains are based on previously published global rating tools.\textsuperscript{23,24} Overall performance was assessed by the summary item ‘Overall ability to perform procedure’, rated on a scale of 1–6, where 1 = not competent to perform independently, 3 = competent to perform independently, and 6 = of above average competence to perform independently.
Development of an error-focused checklist

The error-focused checklist was constructed based on input from an expert panel. To inform items for the survey to administer to this expert panel, two investigators (IWYM, MEB) first independently reviewed a random sample of 16 previously rated performance videos in order to compile a list of serious and common procedural errors. These videos were sourced from 34 available video-recordings of performances of LP recorded at the University of Calgary internal medicine residency programme formative simulation-based examination, which took place between July and September 2011. This examination and its results have been previously described.

Expert panel participants for construction of the error-focused checklist

Once a list of errors had been compiled, 13 members from four Canadian tertiary academic health centres (University of British Columbia, n = 2; University of Calgary, n = 4; University of Toronto, n = 4; University of Ottawa, n = 3) completed three rounds of survey between December 2013 and July 2014. Experts included three haematologists, two neurologists, two internists, one emergency medicine specialist, one anaesthesiologist and four paediatric critical care specialists.

Rounds of survey

The first round of the survey consisted of 38 procedural errors. Experts were asked to rate each error on a 5-point Likert scale based on its likelihood of causing patient harm (1 = not very likely, 5 = very likely), and on a 4-point Likert scale based on the potential consequence of such harm (1 = negligible, 4 = catastrophic). Consensus was defined as agreement of 80% or higher. Experts were also asked to list any additional errors they considered to be clinically significant and to report their procedural experience.

Items that did not achieve consensus (i.e. < 80% agreement) were readdressed in Round 2, in which experts were asked how they would rate the performance (pass versus fail) if the item was the only error witnessed in the performance. A ‘fail’ was to indicate that the trainee was unable to perform the procedure independently, whereas a ‘pass’ was to indicate that the trainee was able to perform the procedure independently. Items that did not reach consensus were readdressed in Round 3. Lastly, experts were asked to rate the relative importance of nine elements to the rating of a trainee’s procedural competence. These elements were: patient safety; comfort; overall success; sterility; time and motion; instrument handling; procedural flow; knowledge of procedure and equipment, and seeking help where appropriate.

Classification of errors

Negligible error

An error was considered negligible if at least 80% of the experts polled in Round 1 agreed that the error was not very likely or somewhat unlikely to cause patient harm and that the harm was likely to be negligible or minor. An error was also considered negligible if at least 80% of experts agreed they would pass the performance if they observed the error.

Serious error

An error was considered serious if at least 80% of the experts in Round 1 agreed that the error was somewhat likely or very likely to cause patient harm and that the harm was serious or catastrophic. An error was considered serious if at least 80% of the experts agreed in Round 2 that they would fail the performance if they observed the error. Results from the three rounds of surveys were then used to create the error-focused checklist.

Rating of performances using the conventionally constructed checklist

All of the remaining 18 video performances that had not been used to inform the error survey items were rated by two independent trained raters using a 21-item conventionally constructed checklist and the eight-item global rating scale as previously described (Appendix S1). Ratings were performed in January 2012 by two trained raters: an internist (IWYM) with 10 years of experience in teaching and assessing procedural skills, and a senior internal medicine resident, who was in her last year of training and had been teaching for 2 years as a certified procedural trainer on the residency training programme. The two raters were trained to consensus for 2 hours on the use of the assessment tools. As described previously, in the rating of the video performances, the order in which the tools were used was alternated with each video in order to minimise the extent to which the rating on one tool might...
systematically influence the rating on the subsequent tool.18

Rating of performances using the error-focused checklist

All 18 video performances were rated by one trained rater using the error-focused checklist, and a random 50% of these videos (i.e. nine videos) were rated independently by a second trained rater in October 2014. The remaining nine videos were rated by the second rater in May 2015. The first rater was the person who had rated the videos in January 2012 using the conventionally constructed tools (IWYM). The second rater (MEB) was a general surgeon with 10 years of experience in teaching procedural skills at both undergraduate and residency training levels. No training on the use of the error-focused checklist was given; however, both raters had extensive experience in the assessment of procedural skills.

Competence/incompetence decisions

Competence/incompetence decisions were based on the summary item on the global rating scale. All performances that achieved a rating of ≥ 3 (competent to perform independently) were considered competent, whereas performances rated as ≤ 2 (borderline competence to perform independently or not competent to perform independently) were considered incompetent. This study was approved by the Conjoint Health Research Ethics Board at the University of Calgary.

Validity evidence

In addition to presenting content validity evidence as outlined above, we assessed for additional sources of validity evidence.25,26 These included internal structure (internal consistency, inter-rater reliability) and relations to other variables (trainees with versus without formal training, checklist scores versus global scores).

Statistical analyses

The sensitivity and specificity of all possible conventional and error-focused checklist cut scores for identifying competence and incompetence, respectively, were evaluated using receiver operating characteristic (ROC) analyses. The area under the curve (AUC) was estimated as a measure of diagnostic accuracy: an AUC of 1.0 indicates perfect diagnostic accuracy.27 Discrimination indices ($D$) were calculated for each conventional checklist item and poorly discriminating items ($D < 0.1$) were removed for the modified conventional checklist scores.28 The AUC of the modified conventional checklist was then re-estimated.

Inter-rater reliability was assessed using intraclass correlation coefficients (ICCs, two-way random model) and kappa statistics. Internal consistency was assessed using Cronbach’s alpha. Scores between two groups were compared using Student’s $t$-tests and correlations between scores were assessed using Pearson’s correlation coefficient. All analyses were performed using PASW Statistics for Windows Version 18.0 (SPSS, Inc., Chicago, IL, USA) and STATA Version 11.0 (StataCorp LP, College Station, TX, USA).

RESULTS

Expert ratings of error items

All 13 experts completed Round 1, and 12 experts (92%) completed Rounds 2 and 3 of the survey. Of the 38 errors considered, 21 items (18 negligible and three serious errors) reached consensus in Round 1 (Table 1). Of the remaining 17 errors, two (items 12 and 17) were felt to be too similar (Table 1) and were collapsed into one item for the remaining two rounds. Two additional errors were suggested by the experts: not performing a Joint Commission Accreditation of Health Care Organizations (JCAHO) time out29 and not using ultrasound in patients with difficult landmarks.30 Thus, in Round 2, a total of 18 errors were surveyed. Of these, 10 items (one negligible and nine serious errors) reached consensus (Table 1). Of the remaining eight errors, five (two negligible and three serious errors) reached consensus in Round 3 (Table 1). No consensus was reached for three items.

From the three rounds of survey, a total of 15 errors were considered serious or deserving of a failure rating, and 21 were considered negligible. Of the 15 serious errors, 13 were applicable to our simulation-based examination (Appendix S2).

Expert panel participants’ experience with lumbar punctures

Ten of the 13 experts (77%) had performed more than 50 LPs, two (15%) had performed between 31 and 40 LPs, and one (8%) had performed between 21 and 30 LPs. Of the 12 experts who provided
Table 1  Error items in expert panel survey and responses from 13 experts

| Item                                                                 | Round in which item reached consensus | Results     |
|----------------------------------------------------------------------|---------------------------------------|-------------|
| 1 Does not wash hands                                                | 3                                     | Serious     |
| 2 Does not position patient appropriately                            | 2                                     | Serious     |
| 3 Does not landmark prior to sterilising and draping                 | 3                                     | Serious     |
| 4 Does not wear mask                                                 | 1                                     | Negligible  |
| 5 Does not wear sterile gloves                                       | 2                                     | Serious     |
| 6 Does not open tray in a sterile manner                             | 1                                     | Serious     |
| 7 Does not open collection tubes ahead of time                       | 1                                     | Negligible  |
| 8 Does not place collection tubes in order                           | 1                                     | Negligible  |
| 9 Places any portion of the manometer outside the sterile field      | 1                                     | Negligible  |
| 10 Cannot properly connect or set up manometer                      | 1                                     | Negligible  |
| 11 Neglects to clean/sterilise the target area altogether            | 2                                     | Serious     |
| 12 Cleans but in so doing, portions of gloved hands touch the patient’s non-sterile back | NA (merged with item 17)              | NA (merged with item 17) |
| 13 Does not allow chlorhexidine to dry in between                    | 1                                     | Negligible  |
| 14 Disposes of used chlorhexidine sponge sticks back into sterile equipment | 2                                     | Serious     |
| 15 Fails to put sterile drape on in a sterile manner (gloves make contact with non-sterile aspect of patient) | 1                                     | Serious     |
| 16 Uses only the fenestrated drape but not the second drape that is normally placed between the patient and the bed | 1                                     | Negligible  |
| 17 Sterile gloves make contact with any non-sterile surfaces (patient, bed, etc.) during the procedure | 2                                     | Serious     |
| 18 Does not warn patient prior to injecting anaesthetic              | 1                                     | Negligible  |
| 19 Does not aspirate prior to injecting local anaesthetic in order to ensure needle not in bloodstream | 1                                     | Negligible  |
| 20 Does not infiltrate deeper tissues with longer needle             | 3                                     | Negligible  |
| 21 Uses lidocaine with epinephrine                                   | 1                                     | Negligible  |
| 22 Does not allow time for anaesthetic to take effect prior to inserting lumbar puncture needle | No consensus                          | No consensus |
| 23 Does not place bevel parallel to nerve fibres                     | No consensus                          | No consensus |
| 24 Stylet not in needle prior to insertion of needle                  | 2                                     | Serious     |
| 25 Inserts needle at a site that is anatomically too high            | 2                                     | Serious     |
| 26 Does not remove stylet completely to check for fluid              | 3                                     | Serious     |
| 27 Fails to check opening pressure                                   | 1                                     | Negligible  |
| 28 Does not ask patient to extend legs prior to measurement of opening pressure | 1                                     | Negligible  |
| 29 Only checks opening pressure after some fluid has already been obtained | 1                                     | Negligible  |
| 30 Does not know how to manoeuvre the stopcock on the manometer     | 1                                     | Negligible  |
| 31 Collects too little fluid per tube                                | No consensus                          | No consensus |
| 32 Collects excessive amount of fluid                                | 3                                     | Negligible  |
| 33 Does not screw on caps of tubes right after fluid collection (but does so at the end) | 1                                     | Negligible  |
| 34 Does not screw on caps of tubes at all                           | 2                                     | Serious     |
| 35 Tries to aspirate cerebrospinal fluid out of canal                | 1                                     | Serious     |
| 36 Does not place stylet back in prior to withdrawing needle         | 1                                     | Negligible  |
| 37 Does not place bandage over site                                  | 1                                     | Negligible  |
information on supervisory experience, eight (67%) had supervised more than 50 LPs and four (33%) had supervised between 41 and 50 LPs.

**Expert ratings of the importance of elements of the procedure**

Overall, experts rated procedural safety, sterility and seeking help as the most important elements in determining procedural competence. Instrument handling, time and motion, and procedural flow were rated as less important (Table 2).

**Performance scores**

Mean ± standard deviation (SD) scores for the 18 videotaped performances were 77.7 ± 14.7% using the conventionally constructed tool and 1.8 ± 1.4 using the error-focused checklist, on which each error is given 1 point for a maximum of 13 points. A higher score on the error-focused checklist indicates a poor performance, whereas the reverse is true for the conventionally constructed tool. A median of two errors per performance was observed (interquartile range [IQR]: 1–3; range: 0–5). The error committed most frequently by trainees was touching non-sterile surfaces with sterile gloves (n = 7, 39%) (Table 3).

Based on overall global rating scale scores, the performances of four (22%) participants were considered competent and the performances of 14 (78%) participants were considered incompetent.

**Accuracy of the assessment tool**

The accuracy of the conventional checklist in identifying competence was high (AUC 0.89, 95% confidence interval [CI] 0.72–1.00) in comparison with that of the error-focused checklist (AUC 0.15, 95% CI 0.00–0.33).

In the identification of incompetence, the accuracy of the conventional checklist was poor (AUC 0.11, 95% CI 0.00–0.28), whereas that of the error-focused checklist was high (AUC 0.85, 95% CI 0.67–1.00).

Overall, conventional checklist cut points demonstrated low specificities for the identification of incompetence, whereas error-focused checklist cut scores demonstrated higher specificities (Table 4).

---

**Table 1 (Continued)**

| Item | Round in which item reached consensus | Results |
|------|--------------------------------------|---------|
| 38 | Places patient on bedrest post-procedure | 1 | Negligible |
| 39 | Does not perform JCAHO-recommended ‘time out’ to verify patient’s identity/verify the type of procedure being done/verify that consent is obtained/verify site* | 2 | Serious |
| 40 | Does not use ultrasound to assess patients in whom landmarks cannot be appreciated* | 2 | Negligible |

JCAHO = Joint Commission on Accreditation of Health Care Organizations; NA = not applicable.
* New items proposed for Round 2 and not included in Round 1 of survey.

**Table 2 Importance of procedural elements in determining how experts rated trainees**

| Element | Score, mean ± SD* |
|---------|-------------------|
| Patient safety | 4.9 ± 0.3 |
| Sterility | 4.8 ± 0.6 |
| Seeking help where appropriate | 4.7 ± 0.5 |
| Patient comfort | 4.4 ± 0.5 |
| Overall success (e.g. obtained CSF) | 4.1 ± 0.7 |
| Knowledge of procedure and equipment (e.g. obviously familiar) | 4.1 ± 0.9 |
| Flow of procedure and forward planning (e.g. effortless flow) | 3.6 ± 0.7 |
| Time and motion (e.g. maximum efficiency) | 3.4 ± 0.7 |
| Instrument handling (e.g. fluid movement, no awkwardness) | 3.1 ± 0.9 |

CSF = cerebrospinal fluid; SD = standard deviation.
* 1 = very unimportant, 5 = very important.
Using the conventional checklist, all competent performances were scored at $\geq 85\%$ (100% sensitivity) and a cut score of $> 97.6\%$ was required to identify incompetence at 100% specificity. Using the error-focused checklist, all performances deemed competent demonstrated no errors and the occurrence of two errors or more was able to identify incompetence at 100% specificity.

**Modified checklist score**

The mean $\pm$ SD discrimination index for the conventional checklist was $0.18 \pm 0.11$. Four items with a discrimination index of $< 0.1$ were removed (‘Withdraw anaesthetic with syringe’; ‘Place sponge stick into chlorhexidine and clean the skin twice in a circular motion from the target area’; ‘Place sterile drape between hip and bed’; ‘Place bandage over puncture site’), resulting in a modified 17-item checklist. The 17-item checklist score highly correlated with the original 21-item checklist score ($r = 0.98$, $p < 0.0001$). The modified checklist had similarly high accuracy in its identification of competence (AUC 0.91, 95% CI 0.78–1.00) and low accuracy in its ability to identify incompetence (AUC 0.09, 95% CI 0.00–0.22).

Table 3  Frequency of errors committed and global scores of 18 participants

| Checklist error items                                                                 | Participants committing error, $n$ (%)* |
|---------------------------------------------------------------------------------------|----------------------------------------|
| Does not wash hands                                                                   | 6 (35)                                 |
| Does not landmark prior to sterilising and draping                                    | 2 (11)                                 |
| Does not open tray in a sterile manner                                                | 2 (14)                                 |
| Does not wear sterile gloves                                                         | 1 (6)                                  |
| Does not clean/sterilise the target area twice with chlorhexidine in a circular motion from the target area | 3 (17)                                 |
| Disposes of used chlorhexidine sponge sticks back into sterile equipment              | 3 (18)                                 |
| Fails to put sterile drape on in a sterile manner (gloves make contact with non-sterile aspect of patient) | 2 (11)                                 |
| Sterile gloves make contact with any non-sterile surfaces (patient, bed, etc.) during the procedure | 7 (39)                                 |
| Inserts needle at a site that is anatomically too high                                | 0                                      |
| Stylet not in needle prior to insertion of needle                                     | 0                                      |
| Does not remove stylet completely to check for fluid                                 | 1 (6)                                  |
| Tries to aspirate cerebrospinal fluid out of canal                                   | 1 (6)                                  |
| Post-collection, does not screw on caps of tubes at all                               | 4 (33)                                 |

| Global rating scores                                                                 | Mean $\pm$ SD                          |
|--------------------------------------------------------------------------------------|----------------------------------------|
| Appropriate preparation of instruments pre-procedure†                               | 2.3 $\pm$ 1.2                          |
| Appropriate analgesia†                                                               | 2.7 $\pm$ 0.8                          |
| Time and motion‡                                                                    | 2.6 $\pm$ 1.1                          |
| Instrument handling‡                                                                | 2.8 $\pm$ 1.1                          |
| Flow of procedure and forward planning‡                                              | 2.6 $\pm$ 1.4                          |
| Knowledge of instruments‡                                                            | 2.5 $\pm$ 1.3                          |
| Aseptic technique‡                                                                  | 2.2 $\pm$ 1.0                          |
| Seeks help where appropriate‡                                                       | 2.5 $\pm$ 1.0                          |
| Overall ability to perform procedure‡                                                | 1.8 $\pm$ 1.1                          |

* Denominator not consistently 18 as some items are either missing or non-applicable (e.g. missing video section, $n = 1$; unsuccessful at obtaining cerebrospinal fluid, $n = 5$; tray already opened, $n = 4$; no sterilisation attempted, $n = 1$).

† Items rated on a scale of 1–6, where 1 = not competent to perform independently and 6 = of above average competence to perform independently.

‡ Items rated on a scale of 1–5, where 1 = not competent to perform independently and 5 = of above average competence to perform independently.
Additional sources of validity evidence

Internal structure

The internal consistency of the error-focused checklist was 0.35, lower than that reported for our conventionally constructed checklist (0.79). The internal consistency of the modified 17-item checklist was 0.78. The internal consistency of the eight-item global rating scale was 0.79. Inter-rater reliability for both the conventional checklist and error-focused checklist was high (conventional checklist: ICC 0.99, 95% CI 0.98–1.00; error-focused checklist: ICC 0.89, 95% CI 0.69–0.96). Inter-rater reliability for the summary global rating score was also high (ICC 0.87, 95% CI 0.73–0.94) and there was perfect agreement between the two raters on determining competence versus incompetence (κ = 1.00).

Relations to other variables

Scores on both the conventional and the error-focused checklists were highly correlated with the summary global rating (conventional checklist: r = 0.61, p = 0.01; error-focused checklist: r = −0.64, p = 0.004). Conventional checklist scores did not differ significantly between trainees who reported having formal LP training (77.1 ± 0.14%) and those who reported no formal training (73.8 ± 0.18) (p = 0.72). Trainees who reported formal training demonstrated fewer errors (1.73 ± 1.1) compared with those without formal training (2.25 ± 2.2), but the difference was also not significant (p = 0.54).

DISCUSSION

Experts in our study considered patient safety, sterility and the trainee’s ability to seek help where appropriate as the most important elements in the rating of procedural competence. Elements such as procedural flow, time and motion, and instrument handling were rated by our experts as being less important. As is consistent with these views on the importance of patient safety and sterility, of the two checklists used in this study, that composed of items referring to errors considered to be serious in nature by our experts demonstrated higher specificity for the identification of ‘incompetence’ than the conventionally constructed checklist. Thus, the diagnostic ability of the error-focused checklist in identifying incompetence was superior to that of the conventionally constructed checklist, whereas the conventionally constructed checklist was superior at identifying procedural competence.

Together, these results make an argument for tailoring the assessment tool to the purpose of the assessment. If the purpose of the assessment is to identify individuals who are incompetent at the procedure (i.e. require more training prior to performing the procedure clinically in patients), the use of an error-focused tool may be preferred. In the era of competency-based education, our study results suggest that the choice of assessment tool may impact on the determination of competence versus incompetence. In our sample, the presence of two or more errors based on the error-focused checklist was uniformly associated with incompetence in performing the procedure and the diagnostic accuracy of the tool for identifying incompetence was high. Although the conventional checklist demonstrated high diagnostic accuracy for competence, its accuracy for determining incompetence was limited.
Thus, in order to identify performances that indicate incompetence in the procedure, a very high cut score on the conventional checklist is required (97% in our sample). However, at this cut score, all but one individual would have been deemed incompetent. Hence the utility of the conventional checklist in determining incompetence may be limited. Even after eliminating poorly discriminating items from our conventional checklist, we were unable to improve the conventional checklist’s ability to detect incompetence.

Assessment of procedural competence should not be a one-time event. Rather, any assessments should be situated within a larger framework of a programme of assessment. For the purposes of formative assessment, early on in a trainee’s procedural education, an educator may prioritise maximising the ability to identify performances that are incompetent and that may pose serious safety risks to patients. Such trainees may benefit from additional training or assessments. The present results suggest that the use of the error-focused tool may be preferable in identifying these trainees. For trainees whose performances were not grossly incompetent based on the error-focused tool, future assessments may then benefit from the use of the conventional tool, which has greater ability to identify competence. However, it is important to note that our study was not designed to assess the conventional tool’s ability to accurately identify competence at the high-stakes level. In this context, repeated formative assessments are recommended.

Unlike a previous study on using clinically discriminating checklists for non-procedural skills, our error-focused checklist did not demonstrate inter-rater reliability or internal consistency higher than those of our conventionally constructed checklist. Although the inter-rater reliability of the error-focused checklist was reasonable (ICC 0.92), its internal consistency was low, which may reflect the fact that procedural incompetence is not a unidimensional concept. Our study sample is too small to further explore its dimensions. In addition, the implementation of error-focused tools will require further study. For example, although one of the errors (Does not screw caps on collection tubes) was deemed serious in nature by our expert panel, the raters did not flag the two performances in which this was the sole error as incompetent. Secondly, two performances without errors were rated as showing ‘borderline competence’. In one performance, seven attempts were made. The second performance, despite the lack of an observed error, was rated as ‘borderline’ because the trainee showed incomplete knowledge of the equipment. None of these procedural issues was appropriately captured in our current error-focused checklist. Thirdly, in procedural assessments, timing matters. At present, both checklists may benefit from additional specificity for the ratings of the items. For example, for an item such as ‘Washes hands’ (or the error-focused equivalent: ‘Does not wash hands’), we rely on the raters to exercise their judgement in determining how to rate trainees who do wash their hands, but do so late into the procedure. Our assessment tools do not capture fully the complexities of the judgement required to appropriately rate all items.

Our study has a number of additional limitations that impact on the interpretation of our results. Firstly, our sample size is small and our study is a single-centre study on one procedure, rated by only two experts. The generalisability of our conclusions to other centres, procedures or non-procedural skills, or to ratings by non-experts, may be limited. Second, despite the use of a panel of experts sourced from across the nation, with multiple specialty representation, the items generated are not necessarily evidence-based items. Further, perhaps due to the automated nature of their expertise, experts may paradoxically neglect key elements of a procedure. Therefore, the sole reliance on experts may be problematic. Thirdly, as indicated earlier, internal consistency in our error-focused checklist was low ($\alpha = 0.35$), which may be a function of having fewer items on the checklist or may indicate that incompetence may be a multidimensional construct. After all, there are likely to be multiple ways in which one can be incompetent. Our sample size is too small to further explore this, but future studies should consider exploring the number of dimensions captured by errors. Fourthly, the purpose of our study and, consequently, the items on our surveys were focused exclusively on procedural errors, their likelihood of causing harm, and the anticipated severity of patient harm. It is highly likely that the framing of these questions would have biased experts into declaring that safety parameters are the most important elements in the rating of procedural competence. It is therefore important to highlight that this bias towards patient safety is prominent in this research. Fifthly, although participants who reported prior formal training achieved higher checklist scores and committed fewer errors than those who reported no formal training, the differences in scores and errors were not statistically sig-
nificant. This lack of difference may reflect any one or more of the following: small sample size; imperfect reliabilities in score measurements; insufficient training in those who received formal training, and learning through clinical exposure in those who did not receive formal training. Sixthly, some items for rating were missing or not applicable as a result of procedural issues. For example, despite our standardised instructions to the examiners to repack the procedural tray at the procedural examination between candidates, the procedural tray was already open at the beginning of the station in four cases. Therefore, those participants did not have an opportunity to demonstrate their technique in tray opening. Further, one video was not started in time and therefore did not capture the participant’s entry to the examination room. Raters were therefore unable to assess whether or not that participant had washed his or her hands. We were unable to rate the ability to screw on the caps of the tubes in candidates who were unsuccessful in obtaining cerebrospinal fluid. Similarly, we were unable to evaluate the disposal of chlorhexidine sponge sticks in candidates who neglected to clean or sterilise the patient altogether. Lastly, we did not explore the use of a combined tool (including both conventional items and error-focused items) and nor did we assess for the acceptability of use of the error-focused tool. For example, if faculties strongly dislike the error-focused tool, lack of acceptability will pose a significant barrier to its use.35 Future studies should address these gaps.

Despite our study’s limitations, our results do suggest that modifying the type of items included in a procedural checklist, specifically in this study by including items on procedural errors, can enhance its ability to detect procedural incompetence. The use of an error-focused checklist identified fewer false positives (i.e. those who are incompetent, but mislabelled as competent) than simply setting a very high cut score on a conventionally constructed checklist. The use of an error-focused checklist should be considered for the determination of incompetence. Secondly, this study presents, for the first time, a list of procedural errors in LP that are considered unacceptable from a safety point of view based on the opinion of experts. These errors should be considered in the training of learners in performing the procedure and may provide guidance to novice faculty raters tasked with assessing procedural competence.

It is worth noting that our conclusions are hypothesis-generating and pertain primarily to the LP procedure, in which procedural errors may cause significant patient harm. These results should not be extrapolated to other clinical skills, especially those in which the clinical consequences of errors are less clear. In these instances, conventionally constructed checklists may be preferred. Future studies should assess whether or not the additional modification of conventional checklist items, such as by weighting items, may be able to yield a more valid tool. If not, the superiority of error-focused tools in the identification of incompetence needs to be confirmed for other procedural skills, and additional sources of evidence of validity, such as the response process and consequences of testing, should be examined. Lastly, the role of errors in the teaching of procedural skills should also be explored.

Contributors: IWYM contributed to the conception and design of the study, and the acquisition, analysis and interpretation of data, and drafted the paper. DP and BM contributed to the conception and design of the study, and the acquisition, analysis and interpretation of data. MEB contributed to the conception and design of the study, and the analysis and interpretation of data. LC and JNS contributed to the acquisition, analysis and interpretation of data. All authors contributed to the critical revision of the paper and approved the final manuscript for publication.

Acknowledgements: the authors wish to thank the experts who participated in the expert panel for this study.

Funding: this study was funded by a Royal College of Physicians and Surgeons of Canada Medical Education Research Grant. The funder had no role in the design and conduct of the study, the collection, management, analysis and interpretation of data, the preparation, review or approval of the manuscript, or the decision to submit the manuscript for publication.

Conflicts of interest: none.

Ethical approval: this study was approved by the Conjoint Health Research Ethics Board at the University of Calgary (Ethics ID E25052).

REFERENCES

1 Accreditation Council for Graduate Medical Education. ACGME programme requirements for graduate medical education in internal medicine. 2009. http://www.acgme.org/acgmeweb/Portals/0/PFAssets/2013-PR-FAQ-PIF/140_internal_medicine_07012013.pdf. [Accessed 26 May 2015.]

2 Accreditation Council for Graduate Medical Education. ACGME programme requirements for graduate medical education in critical care medicine. https://www.acgme.org/acgmeweb/Portals/0/PFAssets/2013-PR-FAQ-PIF/142_critical_care_int_med_07132013.pdf. [Accessed 26 May 2015.]
3 Accreditation Council for Graduate Medical Education. ACGME programme requirements for graduate medical education in emergency medicine. 2013. https://www.acgme.org/acgmeweb/Ports/0/PFAssets/2013-PR-FAQ-PF/110_emergency_medicine_07012013.pdf. [Accessed 26 May 2015.]

4 Royal College of Physicians and Surgeons of Canada. Objectives of training in the specialty of internal medicine. 2011. http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000910.pdf. [Accessed 27 May 2015.]

5 Royal College of Physicians and Surgeons of Canada. Objectives of training in the specialty of general surgery. 2010. http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000902.pdf. [Accessed 27 May 2015.]

6 Royal College of Physicians and Surgeons of Canada. Objectives of training in paediatrics. 2008. http://www.royalcollege.ca/cs/groups/public/documents/document/y2vk/mdaw/~edisp/tztest3rcpsced000911.pdf. [Accessed 27 May 2015.]

7 Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ 2015;49:161–73.

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

9 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.

10 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.

11 Yudkowsky R, Tumuluru S, Casey P, Herlich N, Ledonne C. A patient safety approach to setting pass/fail standards for basic procedural skills checklists. Simul Healthc 2014;9:277–82.

12 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.

13 Clayton MJ. Delphi: a technique to harness expert opinion for critical decision-making tasks in education. Educ Psychol 1997;17:373–86.

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

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

16 Faraggia D, Reiser B. Estimation of the area under the ROC curve. Stat Med 2002;21:3093–106.

17 Hopkins KD. Educational and Psychological Measurement and Evaluation. Needham Heights, MA: Allyn & Bacon 1998.

18 Joint Commission. The universal protocol for preventing wrong site, wrong procedure, and wrong person surgery. Guidance for health care professionals. http://www.jointcommission.org/standards_information/up.aspx. [Accessed 26 May 2015.]

19 Nomura JT, Leech SJ, Shenbagamurthi S, Sierzynski PR, O’Connor RE, Bollinger M, Humphrey M, Gukhool JA. A randomised controlled trial of ultrasound-assisted lumbar puncture. J Ultrasound 2007;26:1341–8.
31 Frank JR, Snell LS, ten Cate O et al. Competency-based medical education: theory to practice. Med Teach 2010;32:638–45.

32 Dijkstra J, van der Vleuten C, Schuwirth L. A new framework for designing programmes of assessment. Adv Health Sci Educ Theory Pract 2010;15:379–93.

33 Lew SR, Page GG, Schuwirth LWT, Baron-Maldonado M, Lescop JM, Paget NS, Southgate JL, Wade DB. Procedures for establishing defensible programmes for assessing practice performance. Med Educ 2002;36:936–41.

34 Sullivan ME, Yates KA, Inaba K, Lam L, Clark RE. The use of cognitive task analysis to reveal the instructional limitations of experts in the teaching of procedural skills. Acad Med 2014;89:811–6.

35 van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract 1996;1:41–67.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Conventional checklist and global rating scale.

Appendix S2. Error-focused 13-item checklist.

Received 10 February 2015; editorial comments to author 11 May 2015, accepted for publication 3 June 2015