BMJ Open Respiratory Research

Ensuring basic competency in chest tube insertion using a simulated scenario: an international validation study

Peter Hertz,[1,2] Katrine Jensen,[1,2] Saleh N Abudaff,[3] Michael Strøm,[1,4] Yousif Subhi,[1] Hani Lababidi,[3] Lars Konge[1]

To cite: Hertz P, Jensen K, Abudaff SN, *et al*. Ensuring basic competency in chest tube insertion using a simulated scenario: an international validation study. *BMJ Open Resp Res* 2018;**5**:e000362. doi:10.1136/bmjresp-2018-000362

Received 20 September 2018
Revised 7 November 2018
Accepted 9 November 2018

[1]Copenhagen Academy for Medical Education and Simulation (CAMES), University of Copenhagen and the Capital Region of Denmark, Rigshospitalet, Copenhagen, Denmark
[2]Department of Cardiothoracic Surgery, University Hospital of Copenhagen, Rigshospitalet, Copenhagen, Denmark
[3]King Fahad Medical City, Riyadh, Saudi Arabia
[4]Department of Vascular Surgery, University Hospital of Copenhagen, Rigshospitalet, Copenhagen, Denmark

Correspondence to
Dr Peter Hertz;
hertz.peter@gmail.com

ABSTRACT

Introduction Chest tube insertion can be associated with serious complications. A structured training programme is needed to minimise complications and enhance patient safety. Novices should pass a reliable test with solid evidence of validity before performing the procedure supervised on patients. The aim of this study was to establish a credible pass/fail standard.

Methods We used an established assessment tool the Chest Tube Insertion Competency Test (TUBE-iCOMPT). Validity evidence was explored according to Messick's five sources of validity. Two methods were used to establish a credible pass/fail standard. Contrasting groups' method: 34 doctors (23 novices and 11 experienced surgeons) performed the procedure twice and all procedures were video recorded, edited, blinded and rated by two independent, international raters. Modified Angoff method: seven thoracic surgeons individually determined the scores that defined the pass/fail criteria. The data was gathered in Copenhagen, Denmark and Riyadh, Saudi Arabia.

Results Internal consistency reliability was calculated as Cronbach's alpha to 0.94. The generalisability coefficient with two raters and two procedures was 0.91. Mean scores were 50.7 (SD±13.2) and 74.7 (SD±4.8) for novices and experienced surgeons, respectively ($p<0.001$). The pass/fail score of 62 points resulted in zero false negatives and only three false positives.

Discussion We have gathered valuable additional validity evidence for the assessment tool TUBE-iCOMPT including establishment of a credible pass/fail score. The TUBE-iCOMPT can now be integrated in mastery learning programmes to ensure competency before independent practice.

Key messages

- ▶ Time-based and number-based education are being replaced with competency-based education, but how can we ensure basic competency in a reliable and valid way?
- ▶ We have established a credible pass/fail score for assessing chest tube insertion, using a reliable assessment tool, the Chest Tube Insertion Competency Test.
- ▶ The presented assessment tool and pass/fail standard can be used to implement mastery learning programmes for young doctors prior to their clinical practice. Furthermore, the contemporary validity framework and the standard setting methods used in this article can be used to gather the necessary validity evidence concerning other clinical procedures.

INTRODUCTION

Chest tube insertion is a common procedure that is important to master, as it is often associated with serious complications.[1–4] Several publications state the need for a structured training programme in order to minimise complications such as incorrect anatomical insertion site and extrathoracic tube placement.[2 3 5] Both hospitals and universities have developed training programmes to teach technically difficult invasive procedures, but the effect of these training programmes have until recently only been measured as a higher self-reported confidence after the training[5–7] which does not necessarily correspond with better performance.[8] Some of the programmes have reported a higher skill level after training.[5 6]

Mastery learning in simulation-based medical education is relevant in competency-based training.[9]

The mastery learning concept defines objectives for skill level, thereby ensuring that all trainees will reach a certain level of competence independent of time spent training. This differs from usual courses using a set training time or performance of a certain number of procedures performed; neither of these methods can ensure competence level nor quality of care.[10 11] The Mastery Learning concept requires an assessment tool with solid evidence of validity including a credible pass/fail standard that can be used for passing or failing a trainee. A reliable rating procedure is essential when high stakes assessment for certification purposes are performed.[12 13] Salamonsen *et al*[14] developed an assessment tool, the Chest Tube Insertion Competency Test (TUBE-iCOMPT), to assess competency of chest tube insertion. The TUBE-iCOMPT can be used as an instrument to assess competence in chest tube insertion and guide the instructor in which aspects of the procedure the trainee needs more practice (formative assessment). The authors of the assessment tool explored the reliability and discriminatory ability of the tool, but generalisability to other training environments and raters has not yet been examined. Furthermore, a pass/fail score needs to be established for the TUBE-iCOMPT to set mastery learning criteria that allow the users of TUBE-iCOMPT to determine when a trainee is competent enough to be allowed to proceed to performing the procedure supervised on patients (summative assessment).

The aim of this study was to gather additional validity evidence in an international setting and to establish a credible pass/fail score for the TUBE-iCOMPT when using blunt dissection technique.

METHODS

We used the internationally recommended validity framework described by Messick including five sources of evidence: content, response process, internal structure, relations to other variables and consequences.[15 16] Data was gathered at two medical education centres: Copenhagen Academy for Medical Education and Simulation, Copenhagen, Denmark (DK)[17] and King Fahad Medical City, Riyadh, Kingdom of Saudi Arabia (KSA).[18]

Participants

We included two groups representing novice and experienced chest tube operators. Criteria for novices were: newly graduated doctors, who had never inserted a chest tube. Criteria for experienced were: physicians having inserted at least 30 chest tubes within the last 12 months, using the blunt dissection technique. Participants were recruited from the hospitals: King Fahad Medical City (KFMC) in Riyadh, KSA and Rigshospitalet, Copenhagen, DK. All participants participated voluntarily and all provided written informed consent.

Procedures

Prior to the procedures, the participants were supplied with a 15 min instructional video[19] and 21 slides from a 'How to insert a chest drain' guide.[20] Written instructions were given to all participants to minimise threats to validity related to the *response process*. The procedures were conducted in a standardised simulated clinical setup using a mannequin (Chest Drain and Needle Decompression Trainer, Limbs and Things, Bristol, UK). The mannequin presented anatomical landmarks for finding correct insertion site with palpable ribs on both sides. New chest tube insertion pads were used for every insertion. These contained tissue-like foam, making it possible to conduct realistic blunt dissection and puncture of the pleura. Each participant completed two chest tube insertion procedures with a size 28 chest tube: one on the left side and one on the right side (figure 1). All procedures were video recorded from two angles: one overview and one zoomed in at the insertion site for later video-based rating. The same facilitator was present during all procedures in both countries and no prompting was given. The setup at the two centres was identical.

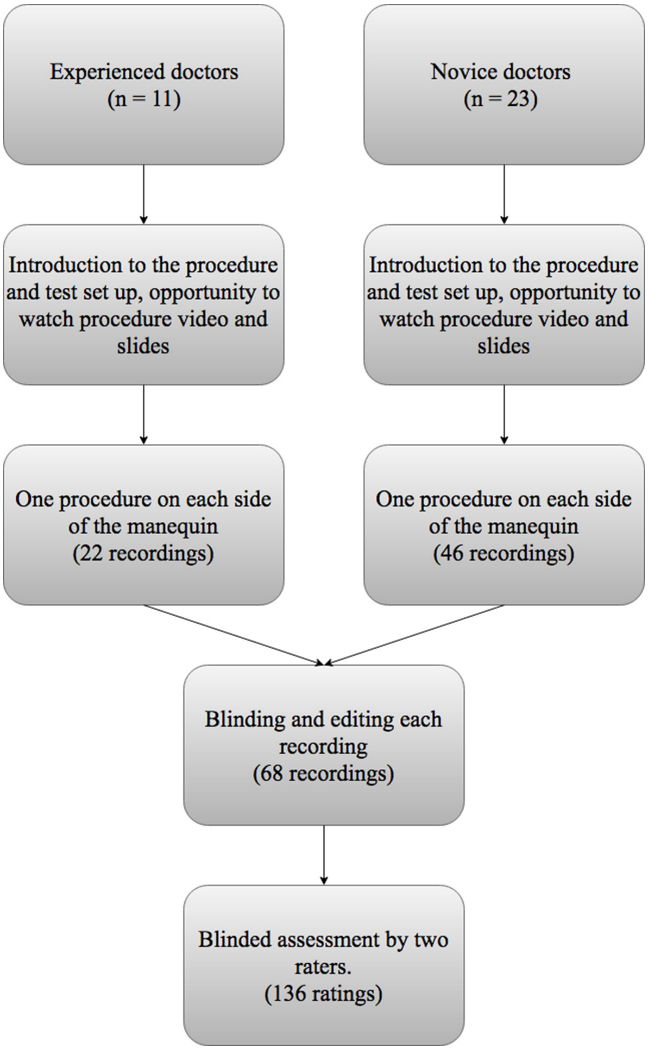
Experienced doctors
(n = 11)
Novice doctors
(n = 23)
Introduction to the procedure and test set up, opportunity to watch procedure video and slides
Introduction to the procedure and test set up, opportunity to watch procedure video and slides
One procedure on each side of the manequin
(22 recordings)
One procedure on each side of the manequin
(46 recordings)
Blinding and editing each recording
(68 recordings)
Blinded assessment by two raters.
(136 ratings)

Figure 1 Flowchart of the participants and rating.

Blinding

Participants were anonymised by wearing a surgical cap, mask and gown. Each video was later edited using *Wondershare Video Editor* (Wondershare Europe, Atena, Germany). The zoomed angle recording was inserted as a picture in picture, covering the participant's head (figure 2).

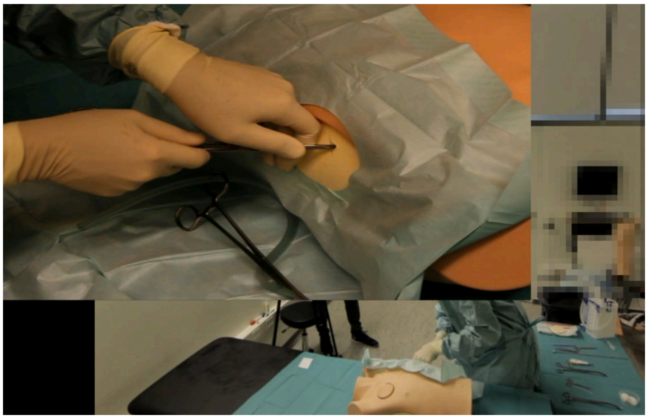

Figure 2 Screenshot from one of the videos.

The assessment tool

Validity evidence concerning *Content* was ensured in the previous study by Salamonsen *et al*[14] in the development of the TUBE-iCOMPT.

The TUBE-iCOMPT assessment tool consists of five domains. The first domain, 'preprocedural checks', was not able to discriminate between experience levels in the original study, and we omitted this domain which left 84 obtainable points.

The TUBE-iCOMPT has two legs and can assess both the Seldinger and the blunt dissection techniques; however we have only investigated the pass/fail standard of the blunt dissection leg.

Raters

The two expert raters were thoracic surgeons, one from DK and one from KSA. The raters did not know each other and did not have any contact during the rating of the videos. The principal author was available in case of technical questions. The raters were given the TUBE-iCOMPT and chest drain insertion guidelines[21] and a short written rating guideline to ensure uniform understanding of the rating items. Furthermore, three practice ratings were conducted and the results were compared to clarify major rating differences; only small adjustments were needed.

Scoring

The edited and blinded video recordings were distributed to the two expert raters by a web-based rating programme[22] showing the video and the assessment tool in the same window. The raters had the possibility to pause and replay the video while rating.

Statistical analysis

As the number of participants in each group was above 10 and since the results are based on distribution of means, it is possible to assume the data as normally distributed.[23]

Internal consistency was investigated using Cronbach's alpha and generalisability theory.[23] Generalisability theory allows exploration of the various types of variance influencing the results.

A decision study (D-study) investigated how many raters and procedures were necessary to ensure reliable test results. A generalisability coefficient above 0.8 is recommended for high stakes assessments.[23]

Independent samples t-test was conducted on the mean scores of each group to explore *relations to other variables*, that is, the experience level of the participants.

Consequences were explored by establishing a pass/fail standard using two different standard setting methods: the contrasting groups' method and the modified Angoff method. In the contrasting groups' method, the pass/fail score is defined by the intersection of a distribution plot of the two groups' mean scores.[24 25] The modified Angoff method uses experts that individually set the score that they believe indicates competence. The experts in the Angoff method were consultant thoracic surgeons. Consultants from each of the four Danish University Hospitals and from King Fahad Medical City were invited to participate. The experts were asked to set the pass/fail criteria to allow a fictional trainee to pass if he or she performed just good enough to proceed to perform the procedure supervised on real patients. Each expert was given oral and written instructions on the method and on how to set the pass/fail score. The pass/fail score was determined as the mean of the experts' contribution.

P values below 0.05 were considered statistically significant.

SPSS V.22 and G-string IV statistical software package (Papaworx, Hamilton, ON Canada) were used for statistical analysis.

RESULTS

Thirty-five participants were included. One was excluded due to a technical error with the video recording. The participants were 23 novices (DK=11, KSA=12) and 11 experienced physicians (DK=6, KSA=5), this leading to a total of 136 completed assessment forms (figure 1).

Validity evidence concerning *content* was ensured in the previous study by Salamonsen *et al*[14] in the development of the TUBE-iCOMPT.

The following actions were taken in order to minimise threats to validity related to the *response process*. Written instructions were given to all participants; the setup and the facilitator were identical for all procedures; procedures were video recorded to allow blinded and independent ratings, and raters were trained using test videos and standardised instructions.

Internal structure was explored by calculating the internal consistency reliability as Cronbach's alpha=0.94. The generalisability coefficient with two raters and two procedures was 0.91. Seventy-seven per cent of the relative variance originated from differences among the participants, 3.2% of the variance were derived from variability among the raters (inter-rater reliability) and only 0.5% of the variance derived from variability among

Table 1 Results from the Generalisability analysis with relative contribution of variance

Results from the G-study showing the contribution of each source of variance

Source of variance V	Description	Relative contribution	Interpretation of results
Participant, V_{pa}	Variation among participants	77.0%	Most of the variance comes from the various skill levels among the participants
The procedures, V_{p}	Variability among the two procedures	0.5%	Almost none of the variance comes from the procedures due to the standardised testing setup
Rater, V_{r}	Variability among the raters	3.2%	A small contribution to the variance indicates a high level of agreement among the raters
Interaction between the participant and the procedures, V_{pa*p}	If the participants consistently show a difference in the two procedures	2.8%	The small contribution to the variance from this interaction indicates a small learning by testing effect and a successful blinding
Interaction between the rater and the participant, V_{pa*r}	If a rater assesses a particular participant differently	2.3%	A small variance contribution indicating a successful blinding
Interaction between the rater and the procedure, V_{p*r}	If a rater shows a consistent difference in rating the two procedures	0.0%	No variance contribution indicates a successful blinding
Interaction between participant, rater and procedures, V_{pa*r*p}	The remaining variability	14.2%	Unavoidable, unexplained error

the two procedures (test–retest reliability). The different contributions to variance are shown in table 1.

Two raters and one procedure or one rater and two procedures were needed to reach a generalisability coefficient above 0.8 (figure 3).

The assessment tool was able to discriminate between levels of experience, which delivers validity evidence for *relations to other variables*. The total mean scores were 50.7 (SD ±13.2) and 74.7 (SD ±4.8) for the novices and the experienced, respectively (p<0.001). The mean difference between groups was 24.0 points with a 95% CI ranging from 17.7 to 30.4.

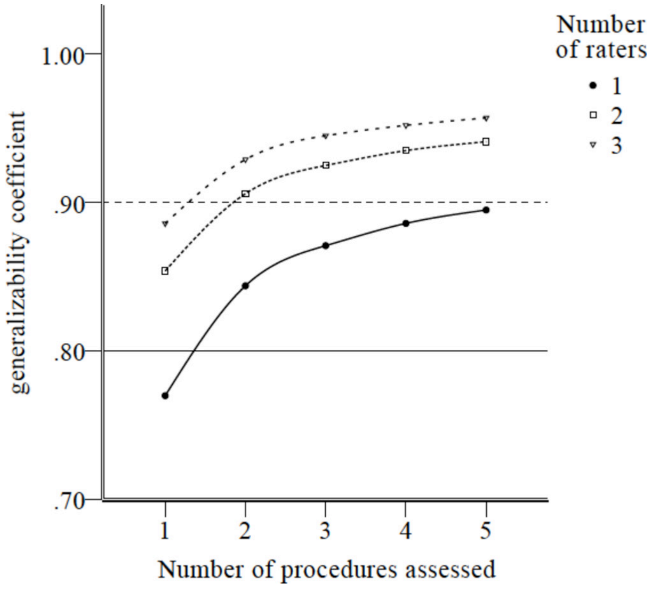

Figure 3 Results from the decision study showing how many raters or procedures are needed in regard to the desired generalisability coefficient.

The pass/fail score established using the contrasting groups' method was 66 points out of 84. Seven consultant thoracic surgeons (five from DK and two from KSA) participated in the modified Angoff sandard setting, and their mean pass/fail score was calculated to 58 (SD ±12.7) (table 2).

Combining the results from the two standard settings gave a pass/fail score at 62 and the *consequences* of the test were zero false negative (experienced who failed the test) and three false positive (novices who passed the test) outcomes.

DISCUSSION

We have gathered additional validity data in an international setting and established a credible pass/fail score for the TUBE-iCOMPT; an existing assessment tool developed for formative assessment.[14] Additional validity evidence according to the recommended contemporary framework for validity[15] was gathered from two international education centres to ensure generalisability of the tool. The contrasting groups' standard setting method and the modified Angoff method (using consultants from five different university hospitals) were used to set a credible pass/fail standard with acceptable consequences. We meet Reznick *et al*'s[26] demands for a large-scale study and generalisable findings across international institutions, making the assessment tool ready for incorporation in competence-based learning programmes with mastery learning criteria.

The context of the content is not changed in this study, as the procedure of chest tube insertion in Australia,

Table 2 Displaying detailed results from each Angoff judge for the separate domains in the Chest Tube Insertion Competency Test

Results from the modified Angoff study								
Angoff judge	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Mean/max (SD)**
Domain 2 score	18	21	25	23	13	17	10	18/28 (5.4)
Domain 3 score	27	26	32	34	25	28	21	28/35 (4.3)
Domain 5 score	13	12	14	19	12	11	7	13/21 (3.6)
Overall score	58	59	71	76	50	56	38	58/84 (12.7)

where the original study originated from, is identical to the procedure in DK and KSA, and in line with British Thoracic Society guidelines.[21]

Several measures were taken to eliminate sources of error in the response process. The data gathering was conducted by the same author; the information and introduction was given in writing in addition to oral to ensure a uniform introduction to all participants. The generalisability analysis (G-study) showed that the contribution of relative variance in relation to response process was low, which indicated uniform setup and successful blinding of the raters. Ma *et al*[27] found no significant differences in direct observation versus blinded video rating of central venous catheterisation, so one could doubt the necessity of video rating and blinding. Contrary to these findings, Konge *et al* found a significant identification bias towards experienced doctors in endoscopic ultrasonography in a study investigating different assessment modalities.[28] We wanted to compare the level of experience and explore the score obtained, using the assessment tool to establish a credible pass/fail score. On this basis, the identification bias is a major threat to validity countered by anonymisation of participants.

Internal consistency reliability of the TUBE-iCOMPT was high with a Cronbach's alpha of 0.94. The generalisability coefficient for our setup was 0.91 and considered very good. In the G-study, we showed that 77% of the relative variance contribution comes from the participants, which is high compared with similar studies.[29] When assessing trainees of different competence levels in a standardised simulated setup, the contributions of variance to the results are important, as we want to measure the true results. Disagreement of the raters only contributed with 3.2% relative variance arguing for a high inter-rater reliability. Only 2.3% of the relative variance originated from the rater–participant interaction (table 1). The D-study showed that two raters and one procedure or two procedures and one rater are sufficient to ensure a generalisability coefficient above 0.8 (figure 3), making the TUBE-iCOMPT feasible for high stakes summative assessment.

Salamonsen *et al*[14] showed that the TUBE-iCOMPT could distinguish participants based on their skill level. The current study supports their findings and shows that the assessment tool can be used by others in an international context, that is, it is generalisable. Leaving out domain one of the original TUBE-iCOMPT did not have any impact on the discriminatory ability which was anticipated from the previous research.[14] In the original study, Salamonsen *et al*[14] found the following mean scores for the intermediates and advanced groups in the blunt dissection 74.3 (95% CI 72.6 to 75.9) and 87.0 (95% CI 85.7 to 88.4), but does not provide data from which domains the points are obtained. Assuming both groups obtained maximum points in domain 1 and by subtracting the 16 points from the omitted domain one in our study, the scores will be 58.3 and 71.0, respectively. With our newly found pass/fail score at 62 points, the groups in the original study will be divided by skill level.

This study was not performed on real patients due to practical and ethical considerations. Instead, we used a mannequin and had the participants perform the procedure in a standardised simulated clinical setting. A relevant concern is the quality of the transfer from the simulated setting to the real patient in the hospital setting. The standardised simulated clinical setting was as lifelike as possible, to make the score obtained in our setting as similar to the hospital setting as possible.[30]

In the original study by Salamonsen *et al*,[14] intermediate and advanced participants were rated when performing the procedure on real patients, and their scores were not significantly different from the ones that the groups obtained when performing the procedure on a mannequin. This indicates transfer of skills, taking into account that the assessment was conducted live, with no blinding and only included a small number of participants. Other studies demonstrate a comparable result from the educational setting to the clinic in various procedures.[31–35] In a systematic review, Dawe *et al* state that under the right circumstances there is transfer of skills from a simulation to a clinical setting.[36] De Gara[37] questions the transferability of skills learnt in a simulated setup and argues that when isolating technical skills for basic training the skills are 'decontextualised'. In our study, the participants performed the entire procedure in one go and had to describe the next steps for the patient case such as chest X-ray, etc. Thus, the TUBE-iCOMPT gave the trainee the possibility to demonstrate the obtained skills in the full procedural context. De Gara[37] also expresses concern for the false sense of security, after successful simulation training. To counter this, an objective pass/fail standard was

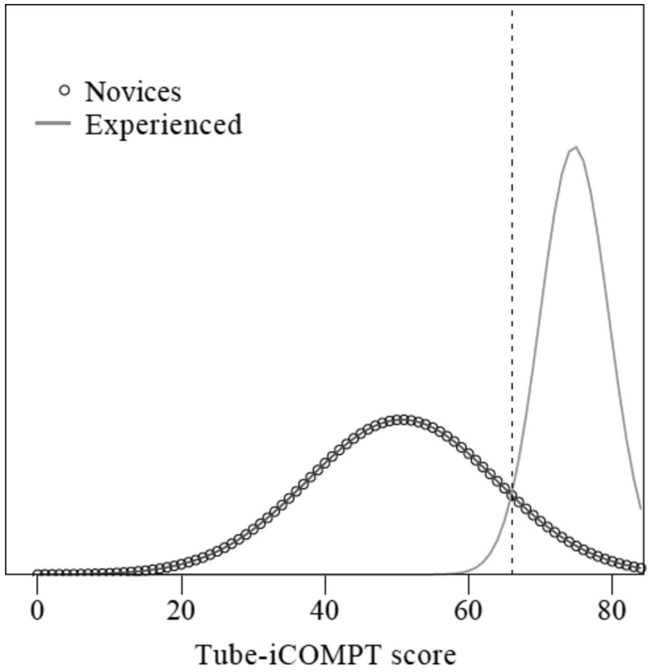

Figure 4 Mean scores of the two groups in the contrasting groups method. The intersection line is the pass/fail score at 66 point.

established for the trainee to progress into bedside learning and supervision in the clinic.

Since there is no gold standard in how to establish a pass/fail standard, the usage of two standard setting methods gave us the ability to find a more accurate and reliable pass/fail score.[24] In the contrasting groups' method, the pass/fail score was found at 66 with zero false negative and three false positives. The pass/fail score is set at the intersection of the two distributions (figure 4) the passing score can be moved left or right to minimise error.[24] Moving the pass/fail score is a policy decision. With respect to the modified Angoff judges whose pass/fail score was found to be considerably lower than the one provided in the contrasting groups' method we did not adjust the pass/fail score in the contrasting groups. Using the pass/fail score from the modified Angoff alone would result in five false positives. Calculating the mean score from multiple standard setting methods has been useful in earlier studies.[38] The final pass/fail score in this study was found by a mean of the two methods' pass/fail scores and resulted in three false positives and zero false negatives.

Data was gathered from two education centres, giving the results an international diversity with participants and raters across nations. The consultants in the Modified Angoff Method have received their thoracic surgery training in different international and national centres, leading to the broad experience in the group of Angoff judges. Our findings and methods contribute to show a high level of generalisability of the TUBE-iCOMPT.

The TUBE-iCOMPT was originally designed with the flexibility to either rate the Seldinger or the blunt dissection technique. Future research is needed to establish a reliable pass/fail standard regarding the Seldinger technique.

CONCLUSION

Additional validity evidence was gathered for the TUBE-iCOMPT as a reliable tool in assessing chest tube insertion skills. A pass/fail score of 62 points out of 84 was established for the blunt dissection technique. It is now feasible and defensible to establish a simulation-based mastery learning training programme in chest tube insertion using the TUBE-iCOMPT to ensure competence before allowing clinical supervised practice.

Acknowledgements The authors would like to thank: Clinical Skills Development Service (Queensland, Australia) (https://www.sdc.qld.edu.au) to licence the use of slides from their course "chest drain course for doctors". Massachusetts Medical Society for permission of the usage of the instruction video "chest tube insertion".

Contributors LK contributed to the development of the statistical analysis plan. HL implemented the data collection in Riyadh. YS was repsonsible for the video data management and statistical analysis. MS contributed to the data collection. SNA and KJ are expert raters. PH was responsible for the data collection, drafted and revised the paper and is the guarantor. LK and PH analysed the data. LK, HL and PH contributed to the design and idea of the study. HL, YS, MS, SNA and KJ contributed to the critical revision of the draft paper.

Funding The two medical education centres covered equipment and travel costs.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Danish Ethical Committee (journal no. 150019066) and Saudi Arabia IRB (reg. IRB00008644 and FWA00018774).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Harris A, O'Driscoll BR, Turkington PM. Survey of major complications of intercostal chest drain insertion in the UK. *Postgrad Med J* 2010;86:68–72.
2. Elsayed H, Roberts R, Emadi M, *et al*. Chest drain insertion is not a harmless procedure – are we doing it safely? *Interact Cardiovasc Thorac Surg* 2010;11:745–8.
3. Maritz D, Wallis L, Hardcastle T. Complications of tube thoracostomy for chest trauma. *S Afr Med J* 2009;99:114–7.
4. Kwiatt M, Tarbox A, Seamon MJ, *et al*. Thoracostomy tubes: a comprehensive review of complications and related topics. *Int J Crit Illn Inj Sci* 2014;4:143–55.
5. Hutton IA, Kenealy H, Wong C. Using simulation models to teach junior doctors how to insert chest tubes: a brief and effective teaching module. *Intern Med J* 2008;38:887–91.
6. Carter YM, Wilson BM, Hall E, *et al*. Multipurpose simulator for technical skill development in thoracic surgery. *J Surg Res* 2010;163:186–91.
7. Sanchez LD, Delapena J, Kelly SP, *et al*. Procedure lab used to improve confidence in the performance of rarely performed procedures. *Eur J Emerg Med* 2006;13:29–31.
8. Barnsley L, Lyon PM, Ralston SJ, *et al*. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Med Educ* 2004;38:358–67.
9. Cook DA, Brydges R, Zendejas B, *et al*. Mastery learning for health professionals using technology-enhanced simulation: a systematic review and meta-analysis. *Acad Med* 2013;88:1178–86.

10. McGaghie WC, Harris IB. Learning theory foundations of simulation-based mastery learning. *Simul Healthc* 2018;13(Suppl 1):1.
11. McGaghie WC, Barsuk JH, Wayne DB. AM last page: mastery learning with deliberate practice in medical education. *Acad Med* 2015;90:1575.
12. Yudkowsky R, Park YS, Lineberry M, *et al*. Setting mastery learning standards. *Acad Med* 2015;90:1495–500.
13. McGaghie WC, Issenberg SB, Barsuk JH, *et al*. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ* 2014;48:375–85.
14. Salamonsen MR, Bashirzadeh F, Ritchie AJ, *et al*. A new instrument to assess physician skill at chest tube insertion: the TUBE-iCOMPT. *Thorax* 2015;70:186–8.
15. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–16.
16. Borgersen NJ, Naur TMH, Sørensen SMD, *et al*. Gathering validity evidence for surgical simulation: a systematic review. *Ann Surg* 2018;267:1063–8.
17. Konge L, Ringsted C, Bjerrum F, *et al*. The simulation centre at Rigshospitalet, Copenhagen, Denmark. *J Surg Educ* 2015;72:362–5.
18. Lababidi H, Munshi F, AlAmar M, *et al*. CRESENT: the Center for Research, Education and Simulation Enhanced Training, King Fahad Medical City, Riyadh, Saudi Arabia. *J Surg Simul* 2015;2.
19. Dev SP, Nascimiento B, Simone C, *et al*. Videos in clinical medicine. Chest-tube insertion. *N Engl J Med* 2007;357:e15.
20. Clinical Skills Development Service. *Chest drain course for doctors Queensland Australia*, 2015.
21. Havelock T, Teoh R, Laws D, *et al*. Pleural procedures and thoracic ultrasound: British Thoracic Society pleural disease guideline 2010. *Thorax* 2010;65(Suppl 2):i61–76.
22. Subhi Y, Todsen T, Konge L. An integrable, web-based solution for easy assessment of video-recorded performances. *Adv Med Educ Pract* 2014;5:103–5.
23. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012;34:960–92.
24. Yudkowsky R. Standard setting. In: Downing SM, Yodkowsky R, eds. *Assessment in health professions education*. New York: Routledge, 2009: 119–48.
25. Jørgensen M, Konge L, Subhi Y. Contrasting groups' standard setting for consequences analysis in validity studies: reporting considerations. *Adv Simul* 2018;3:5.
26. Reznick RK, MacRae H. Teaching surgical skills – changes in the wind. *N Engl J Med* 2006;355:2664–9.
27. Ma IW, Zalunardo N, Brindle ME, *et al*. Notes from the field: direct observation versus rating by videos for the assessment of central venous catheterization skills. *Eval Health Prof* 2015;38:419–22.
28. Konge L, Vilmann P, Clementsen P, *et al*. Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. *Endoscopy* 2012;44:928–33.
29. Todsen T, Tolsgaard MG, Olsen BH, *et al*. Reliable and valid assessment of point-of-care ultrasonography. *Ann Surg* 2015;261:309–15.
30. Alrahbi R, Easton R, Bendinelli C, *et al*. Intercostal catheter insertion: are we really doing well? *ANZ J Surg* 2012;82:392–4.
31. Todsen T, Jensen ML, Tolsgaard MG, *et al*. Transfer from point-of-care Ultrasonography training to diagnostic performance on patients-a randomized controlled trial. *Am J Surg* 2016;211:40–5.
32. Sroka G, Feldman LS, Vassiliou MC, *et al*. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial. *Am J Surg* 2010;199:115–20.
33. 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;254:502–11.
34. McCannel CA, Reed DC, Goldman DR. Ophthalmic surgery simulator training improves resident performance of capsulorhexis in the operating room. *Ophthalmology* 2013;120:2456–61.
35. Barsuk JH, Cohen ER, Potts S, *et al*. Dissemination of a simulation-based mastery learning intervention reduces central line-associated bloodstream infections. *BMJ Qual Saf* 2014;23:749–56.
36. Dawe SR, Pena GN, Windsor JA, *et al*. Systematic review of skills transfer after surgical simulation-based training. *Br J Surg* 2014;101:1063–76.
37. Brindley PG, Jones DB, Grantcharov T, *et al*. Canadian Association of University Surgeons' Annual Symposium. Surgical simulation: the solution to safe training or a promise unfulfilled? *Can J Surg* 2012;55:S200–6.
38. Wayne DB, Barsuk JH, Cohen E, *et al*. Do baseline data influence standard setting for a clinical skills examination? *Acad Med* 2007;82:S105–8.