Surgical Endoscopy (2019) 33:1465–1473
https://doi.org/10.1007/s00464-018-6428-1

Evaluating competency in video-assisted thoracoscopic surgery (VATS) lobectomy performance using a novel assessment tool and virtual reality simulation

Katrine Jensen[1,2,6] · Henrik Jessen Hansen[1] · René Horsleben Petersen[1] · Kirsten Neckelmann[3] · Henrik Vad[4] · Lars Borgbjerg Møller[5] · Jesper Holst Pedersen[1] · Lars Konge[2]

Received: 23 February 2018 / Accepted: 5 September 2018 / Published online: 17 September 2018

Abstract

Background Competency-based training has gained ground in surgical training and with it assessment tools to ensure that training objectives are met. Very few assessment tools are available for evaluating performance in thoracoscopic procedures. Video recordings would provide the possibility of blinded assessment and limited rater bias. This study aimed to provide validity evidence for a newly developed and dedicated tool for assessing competency in Video-Assisted Thoracoscopic Surgery (VATS) lobectomy.

Methods Participants with varying experience with VATS lobectomy were included from different countries. Video recordings from participants' performance of a VATS right upper lobe lobectomy on a virtual reality simulator were rated by three raters using a modified version of a newly developed VATS lobectomy assessment tool (the VATSAT) and analyzed in relation to the unitary framework (content, response process, internal structure, relation to other variables, and consequences of testing).

Results Fifty-three participants performed two consecutive simulated VATS lobectomies on the virtual reality simulator, leaving a total of 106 videos. *Content* established in previously published studies. *Response process* Standardized data collection was ensured by using an instructional element, uniform data collection, a special rating program, and automatic generation of the results to a database. Raters were carefully instructed in using the VATSAT, and tryout ratings were carried out. *Internal structure* Inter-rater reliability was calculated as intra-class correlation coefficients, to 0.91 for average measures ($p < 0.001$). Test/re-test reliability was calculated as Pearson's r of 0.70 ($p < 0.001$). G-coefficient was calculated to be 0.79 with two procedures and three raters. By performing D-theory was found that either three procedures rated by two raters or five procedures rated by one rater were enough to reach an acceptable G-coefficient of ≥ 0.8. *Relation to other variables* Significant differences between groups were found ($p < 0.001$). The participants' VATS lobectomy experience correlated significantly to their VATSAT score ($p = 0.016$). *Consequences of testing* The pass/fail score was found to be 14.9 points by the contrasting groups' method, leaving five false positive (29%) and six false negatives (43%).

Conclusion Validity evidence was provided for the VATSAT according to the unitary framework. The VATSAT provides supervisors and assessors with a procedure-specific assessment tool for evaluating VATS lobectomy performance and aids with the decision of when the trainee is ready for unsupervised performance.

Keywords VATSAT · VATS · Assessment tool · Thoracoscopic lobectomy · Competence evaluation

Objective assessment is a critical aspect of surgical training for providing trainees with feedback and summative assessment [1]. Evaluation provided by logbooks and supervision does not optimally ensure surgical competencies. Competency-based training has therefore gained ground in surgical training, and with it, assessment tools to ensure that training objectives are met since trainees master skills at a different pace [2–4]. A variety of assessment tools for medical education exist in both non-technical and technical skills [5, 6], but very few assessment tools are available for evaluating thoracoscopic procedures [7, 8]. The performance of more

✉ Katrine Jensen
katrine.jensen@rh.regionh.dk

Extended author information available on the last page of the article

and more complex procedures in the operating room has resulted in the development of procedure-specific assessment tools [9–12] instead of generic assessment tools such as the Objective Structured Assessment of Surgical Skills (OSATS). The American Board of surgery has developed procedure-specific assessment tools for general surgery, and the Accreditation Council for Graduate Medical Education (ACGME) has launched the Thoracic Surgery Milestone Project, and hereby acknowledged the need for specialty-specific assessment tools in thoracic surgery [13]. But to use assessment tools for decisions with great implications for the trainee, as passing or failing "a test," the assessment tools need to be thoroughly studied to not question the decisions based on these tools—validity evidence needs to be provided. The currently accepted framework of validity is the unitary framework described by Messick, where validity evidence is provided for five sources (content, response process, internal structure, relation to other variables, and consequences of testing) [14].

The OSATS and similar tools are mostly used for direct observation with the assessor being present at the bench station or in the operating room [15, 16]. But direct observation introduces assessment bias due to the personal relations between trainee and assessor/supervisor [17]. Video recordings of a performed Video-Assisted Thoracoscopic Surgery (VATS) lobectomy provide the possibility of blinded assessment, and they do not require the presence of an experienced assessor at the time of the procedure being performed. By using blinded assessment, rater bias in surgical training can to a great extent be avoided [18], but when assessment tools are used on video recordings, the tool often needs to be modified since skills other than technical skills are difficult to be assessed in video recordings [12].

A procedure-specific tool for assessment of VATS lobectomy competency has recently been developed to provide supervisors with a means for objective assessment either by direct observation in the operating room or from video-filmed performances.

This study aimed to provide validity evidence for this newly developed assessment tool for VATS lobectomy (VATSAT).

Materials and methods

Inclusion

A convenience sample of medical students with no surgical experience, trainees in thoracic surgery with varying experience, and experienced thoracic surgeons in VATS lobectomy were included between September 2015 and January 2016. The novices were recruited among Danish medical students without any surgical experience, and the trainees and experienced surgeons in thoracic surgery were recruited from various thoracic centers from all over the world as part of a training program taking place at the University Hospital of Copenhagen, Rigshospitalet, Denmark, at the cardiothoracic department. Participants were excluded if they had any virtual reality simulation experience.

Ethics

Participants were first informed by KJ of the project and that their data would be anonymous to all other than KJ. They then signed a consent form stating that participation was voluntary and that they at any time and without pressure could withdraw their consent for participation and their data would be deleted. There was no physical or psychological harm to the participants. The assessment of each participant was not discussed with colleagues or other individuals. The participants were not told how they performed compared with others. All participants were assigned a unique trial number, and data were kept under this number. All data were kept strictly confidential on a code-locked hard drive by KJ. No samples from humans were used in the study and no drugs were administered; hence, this study needed no approval from The Danish National Committee on Biomedical Research Ethics, but an inquiry was still sent and returned with protocol no. H-15011006.

Testing

The testing [19, 20] took place at the Copenhagen Centre for Medical Education and Simulation (CAMES), Copenhagen, Denmark [21]. Four identical virtual reality simulators (LapSim®, Surgical Science, Gothenburg) were used. Background data (age, gender, VATS experience) were obtained from the participants. The simulators' built-in instructions of the procedural steps of the simulated scenario of a VATS lobectomy of a right upper lobe were shown to the participants, before they performed a VATS lobectomy of the right upper lobe on the simulator two times in a row. The scenario had to be performed in a certain order, and this was shown at the bottom of the screen in the scenario on the simulator throughout the test. All tests and metrics were recorded on the simulator automatically. For a more detailed description and pictures and videos of the simulator and scenario, please see reference 19 and 20 and http://www.surgicalscience.com.

The VATS lobectomy assessment tool (VATSAT)

The VATSAT was developed especially for assessing competency in VATS lobectomy performance by direct observation in the operating room or from video-filmed performances and is described in detail in a previous publication [22]. The

VATSAT consists of eight items especially developed to rate trainees' VATS lobectomies competencies, and each item can be rated from 1 to 5, where 5 is the best score. Rating anchors are provided at 1, 3, and 5 points.

A modified VATSAT (Fig. 1) was used to rate the videos, leaving five items from the original VATSAT and three excluded since the simulator cannot yet simulate the extent of the tumor, lymph node removal, and removal of the lobe in a bag.

Rating of videos

Each participant produced two video-recorded performances, and each video clip recorded from the simulator was given a unique number and uploaded to an online digital rating system and hereby made anonymous [23]. The unedited videos were blindly rated by three thoracic surgeons with extensive experience in VATS lobectomy using the modified VATSAT. The raters could see the video next

	1	2	3	4	5
1. Dissection of the hilum and veins	Dissection is unsafe and the trainee cannot remove connective tissue from the hilum to identify the veins and prepare them for stapling. Dissection is finally done by supervisor.		Connective tissue and lymph nodes if necessary are removed from the hilum and the veins are identified and prepared for stapling with some hands-on guidance from supervisor.		The hilum is properly and safely dissected and the veins are identified and prepared for stapling without help from supervisor. Trainee checks for single pulmonary vein before stapling.
2. Dissection of the arteries	Dissection is unsafe and the trainee cannot identify the arteries to the affected lobe and prepare them for stapling. Dissection is finally done by supervisor.		The pulmonary artery and arteries for the affected lobe are identified, connective tissue and lymph nodes if necessary are removed and the arteries are prepared for stapling with some hands-on guidance from supervisor.		The pulmonary artery and arteries for the affected lobe are identified, proper and safely dissected and prepared for stapling without help from supervisor. Trainee checks for anatomical variations of the arteries before stapling.
3. Dissection of the bronchus	Dissection is unsafe and the trainee cannot identify the bronchus and prepare it for stapling. Dissection is finally done by supervisor.		The bronchus and bronchial arteries to the affected lobe are identified, connective tissue and lymph nodes if necessary are removed and the bronchus is prepared for stapling with some hands-on guidance from supervisor.		The bronchus and bronchial arteries are proper and safely dissected and prepared for stapling without help from supervisor.
4. Respect for tissue and structures	The trainee uses diathermy/instruments to close to vital structures and tissue (nerves, oesophagus, vessels, lung parenchyma in affected/adjacent lobes) and causes unacceptable inadverte nt damage.		The trainee gently manipulates tissue and vital structures (nerves, oesophagus, vessels, lung parenchyma in affected/adjacent lobes) but occasionally causes inadvertent damage.		The trainee consistently demonstrates appropriate handling of tissue and vital structures with minimal inadvertent damage.
5. Technical skills in general	The trainee handles instruments incorrectly and with too much force, does not keep instruments in the field of vision, is not familiar with most instruments and lacks fluidity and accuracy of hand movements.		The trainee handles instruments adequately, keeps instruments in the field of vision most of the time, is familiar with most instruments but is occasionally stiff and awkward.		The trainee demonstrates complete familiarity with all instruments and handles these correctly and not with too much force, keeps instruments in the field of vision all of the time and has excellent fluidity and accuracy of hand movements.

Fig. 1 The modified VATSAT

to the assessment tool, and pause the video and go back and forth to review if necessary. The videos only showed the output from the screen, thus the surgeons were not seen in the video, and there was no sound.

Outcome measures and data analysis

Outcome measures according to Messick's unitary framework for validity [14]:

Content

The steps taken to ensure that test content reflects the construct it is intended to measure, is described for the virtual reality test in VATS lobectomy and VATSAT in previously published studies (user realism, training tool usefulness, and agreement between experts in the field on design, items, and anchors) [19, 20, 22].

Response process

Describes the test security/quality control of an assessment process to eliminate sources of error to the maximum extent possible. In this study, standardized data collection was ensured by using an instructional element before the test on the simulator and letting the same person perform all data collection [19] and the same three persons rate all videos in a rating program where the video is integrated next to the assessment tool [23], and automatic generation of the results to a database. Raters were carefully instructed on how to rate using the VATSAT, and tryout ratings were carried out and disparities discussed and misunderstandings corrected before the real rating. The VATSAT uses global rating and not checklist ratings, and allows for blinded, unbiased rating.

Internal structure

Explores the assessment tool's psychometric characteristics (the reproducibility and correlation of test items). Intra-class correlation coefficients (single measure and averages measures, absolute agreement definition) were used to explore inter-rater reliability. Test/re-test reliability for the scores from the VATSAT was calculated using Pearson's *r*. Generalizability (G) theory with subsequent D-theory was applied to compare number of procedures needed in combination with number of raters needed to reach a G-coefficient of 0.8.

Relation to other variables

Describe if the assessment score correlates with known measures of competence. Participants will be divided into three groups based on their number of performed real-life VATS lobectomies: Novices: 0 performed procedures, intermediates: 1–49 performed VATS lobectomies, and experienced: ≥ 50 performed VATS lobectomies. The assessment tools' discriminatory ability between novices and experienced and how intermediates scores fit was calculated by using the participants mean scores and using ANOVA with post hoc test. Correlation coefficients for the participants' VATS lobectomy experience and VATSAT scores are given.

Consequences of testing

Here, consequences of the assessment decisions are discussed. The pass/fail score of the assessment tool was established using the contrasting groups' method and explored by reporting the consequences.

IBM SPSS statistics version 23 (IBM, NY, USA) and GENOVA software version 3.1 (Download: http://papaworx.com/Deposit/G_String_Installer.zip) were used for the data analysis, and results with $p < 0.05$ were considered significant.

Results

None of the invited participants had any virtual reality simulator experience, and all performed two procedures, and thus none were excluded, leaving 53 participants with varying experience in VATS lobectomy. The participants came from nine different countries (Australia, Belgium, Denmark, Finland, Germany, Ireland, United Kingdom, Spain, Switzerland). Each participant performed two consecutive simulated VATS lobectomies on the virtual reality simulator, leaving a total of 106 videos. Three raters assessed each video with the modified VATSAT. Background data for the participants and their mean VATSAT scores are shown in Table 1.

Internal structure

Inter-rater reliability was calculated as intra-class correlation coefficients (ICC) to be 0.78 for single measures and 0.91 for average measures with both being highly significant with $p < 0.001$. Test/re-test reliability scores for the participants' first and second procedure (in the following called occasions) were calculated by using the mean score of each video from all three raters yielding a Pearson's coefficient of 0.70 and being highly significant with $p < 0.001$. G-coefficient was calculated to be 0.79 with two occasions and three raters, and the variance of the results is found mainly to be related to the participants and not the raters or occasions (Table 2).

Performing D-theory by altering the number of raters and occasions, it was found that either three occasions rated by two raters, or five occasions rated by one rater were enough

Table 1 Background data and mean VATSAT scores

Group	Participants #	Mean age years	Mean performed VATS lobectomies #	Mean VATSAT score for procedure 1 and 2 ± SD
Novices (0)	17	24	0	12.0 ± 3.4
Intermediates (1–49)	22	39	10	15.5 ± 2.9
Experienced (≥ 50)	14	43	126	17.0 ± 4.2

SD standard deviation

Table 2 The contribution of each source of variance to the VATSAT score

Source of variance	Description	Estimated variance	Relative contribution (%)	Interpretation of results
Participants	Systematic variation among participants	12.1	58	Most of the measured variance are derived from the participants' different VATS experience
Raters	Systematic variation among raters	0.07	0.4	The raters had an equal level of stringency
Procedures	Systematic variation among procedures	0	−0	The procedures were equally difficult
Interaction between participants and procedures	Consistent trend for a participant to perform a procedure differently	4.3	20	Moderate variance were contributed to the participants performance of the procedure
Interaction between participants and raters	Consistent trend for a rater to assess a particular participant differently	1.3	6	There was minimal bias between rater and participant due to effective blinding
Interaction between raters and procedures	Consistent trend for a rater to assess differently on a particular procedure	0.05	0.2	The raters' assessments of the procedures were equal
Interaction between participants, procedures, and raters	All remaining variability	3.2	15	Expected unexplained error

to reach a G-coefficient of 0.8 (Fig. 2), but a practice effect could suppress validity when extrapolating beyond four occasions.

Relation to other variable

The mean total scores for novices and experienced were significant different with $p = 0.001$ using independent samples t test. When comparing all three groups using ANOVA with Bonferroni post hoc test, the mean scores were significantly different between novices and intermediates ($p = 0.009$) but not between intermediates and experienced surgeons. The Pearson correlation coefficient for the participants' VATS lobectomy experience correlated to their VATSAT score was 0.33 with $p = 0.016$ and they are plotted in Fig. 3.

Consequences of testing

The pass/fail score of the assessment tool was established to be 14.9 points by the contrasting groups' method, leaving five false positive (29%) and six false negatives (43%)

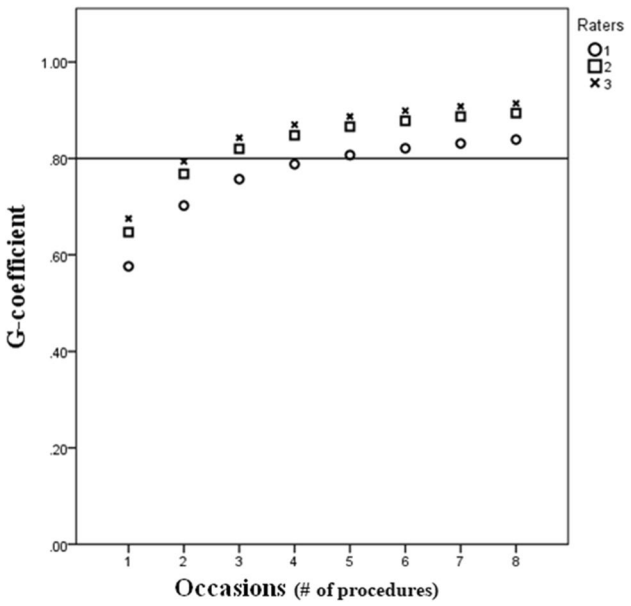

Fig. 2 Number of procedures evaluated by one, two, or three independent raters in order to reach a G-coefficient of 0.8

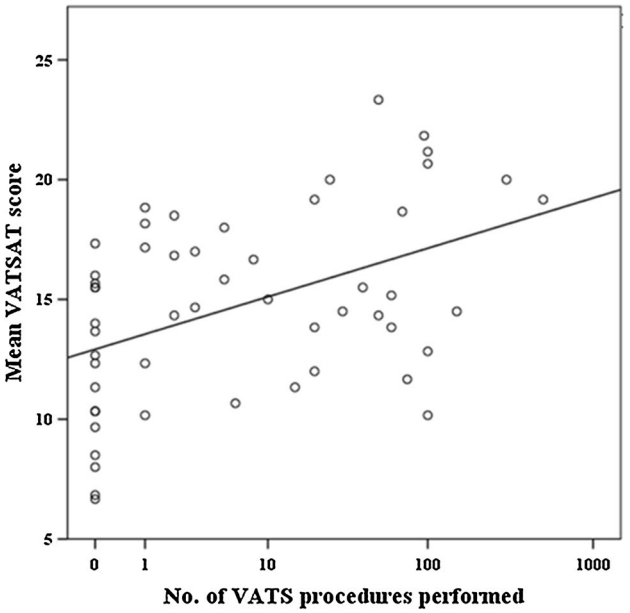

Fig. 3 Mean VATSAT scores in relation to the participants previous VATS lobectomy experience

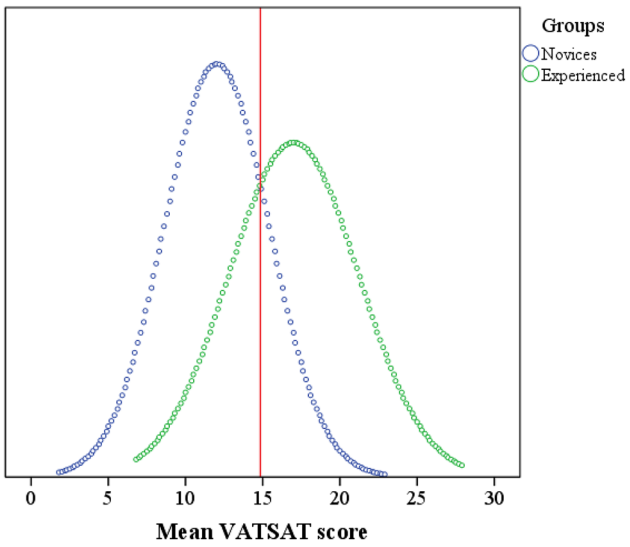

Fig. 4 Pass/fail level using the contrasting groups' method

(Fig. 4). Thirteen of the 22 (59%) intermediate experienced participants scored a mean of more than 14.9 points.

Discussion

In this study, validity evidence was provided for an assessment tool for VATS lobectomy competence. Very high inter-rater reliability with ICC > 0.9 was found for the modified VATSAT. For selection and certification which has major consequences for the surgeon, it is generally agreed that reliability coefficients should be ≥ 0.8 [24]. The tool also had the ability to discriminate in a highly statistically significant way between the defined stages of experience, which improves its usefulness in clinical practice, but as shown in Fig. 4 considerable overlap exists. Pass–fail levels require good validity evidence because of the consequences for the surgeons being tested. The proposed cut score for the VATSAT allowed too many novices to pass and too many experienced to fail, even if it was by a close margin (many passed or failed by half a point). No significant difference was found between intermediates and experienced surgeons, and this may possibly reflect that some of the trainees are quite skilled despite not having performed > 50 VATS lobectomies. The younger surgeons are possibly more intuitive towards the virtual reality simulation than their experienced senior colleagues, and hence some experienced surgeons score low on the VATSAT, but are very skilled surgeons in real life. The current study should be regarded as the initial one in a line of research on the topic, and in this study we used the pass/fail score to distinguish novices from expert and where the intermediates fit with this score. Testing medical students is probably not an appropriate method for validating educational methods aimed for surgical trainees, but this was the first study to explore the VATSAT's discriminatory abilities between levels of expertise, and therefore we wanted to see how well totally untrained medical students scored in comparison with real surgeons. This is important to explore and justify further research into clinical implications of the VATSAT. It requires a clinical study using the full VATSAT before a clinical pass/fail score can be determined and is used to consider a surgeon ready for unsupervised practice. Petersen et al. [25] just published the first results from the clinical application of the VATSAT to real-life VATS lobectomies. Here, the pass/fail score were calculated to 31 point, leaving one of the beginners (not medical students, but surgeons) to pass, and two experts to fail. Bonferroni post hoc tests showed that experts were significantly better than intermediates ($p < 0.008$) and beginners ($p < 0.001$). Intermediates' mean scores were significantly better than beginners ($p < 0.001$).

Generalizability studies use variance component analysis to explore facets that could be a threat to reliability [26]. Rater disagreement only "contributed" to 0.4% of the overall variation in test score (Table 2), which substantiates that a good but short introduction to using the VATSAT was enough. In this study, we found that only 0.2% of the variation in test score was due to the interaction between rater and occasion, hence the raters did not vary in their perceptions of a specific procedure, and this can be related to the fact that the same simulated scenario is equally difficult every time it is performed. Moderate variance was contributed to the participants' performance of the procedure, and this could

reflect that the participants learn from each procedure how to perform smarter/differently in the next attempt.

Virtual reality-simulated VATS lobectomies were used instead of video-recorded real-life surgeries. One could criticize the results based on this, but an advantage of video recordings from a simulator is the standardized scenario, where real patients are very different. Another advantage is that video-filmed performances from a simulator can be acquired for surgeons with different VATS experience (ranging from novices to experienced surgeons). Videos of real-life procedures cannot be obtained from novices (they are not allowed to operate on their own or at all), and intermediate experienced and experienced VATS surgeons could be hesitant about showing their video-recorded procedures (due to possible complications and wanting to show a perfect procedure). Also, the raters would not know if a supervisor was giving the trainee hands-on guidance in the operating room, unless this was registered along with the timestamp on the video. Besides objective observation from video recordings, raters avoid time-consuming attendance in the operating room and have the possibility of re-assessment in case of doubt. Konge et al. [27] compared results based on direct observations and blinded video recordings and found that non-blinded observation favored the consultants. Other studies found that video recordings equalled direct observations [18]. A video-based approach makes it feasible to use the same raters for all procedures, and an important point in using video recordings is the avoidance of all the potential bias of human relations between supervisor/rater and surgeon [17]. However, video rating by blinded raters would be optimal for summative assessment which might incur added costs, but a recent study found that non-specialist raters can provide reliable and valid assessments of video-recorded performance, hereby diminish the rater-cost [28].

However, some important aspects are not possible to be assessed based on video recordings, and simulator performance is never completely realistic and does not reflect all aspects of the clinical procedure. An optimal assessment should be standardized but an optimal training program should be diverse enough to mimic the challenges in real life, and as of now, the simulator cannot simulate how to deal with complications and anatomical variations. Bleeding is simulated, but still not realistic enough for intermediates and experienced surgeons to practice hemostatic control. Simulation training at this developmental stage is primarily for novices to lift them to an intermediate's level before operating on patients, and then real-life procedure training is needed to reach a higher level [29]. The items on the VATSAT focus on technical competency. However, surgical competence includes several other skills, such as cognitive and integrative competencies [30, 31]. Many generic assessment tools for non-technical skills are already developed for this purpose [5], and skilled supervisors using direct observation must ensure the abilities of the trainees in these areas. In the future, maybe wearable video technology can also support this [32, 33].

The strength of this study lies in its design which allows for objectivity and accuracy of results stored in the simulator. Using standards (the simulator) means that the research can be replicated, and then analyzed and compared with similar studies. Raters were carefully instructed and tryout ratings were performed. Data based on participants from many countries increase the generalizability of the results and the sample size was large. The grouping of the participants poses some difficulties, because a very skilled trainee can probably score the same as a very experienced surgeon, and does experience equal good skills/a high score on the test, i.e., does a surgeon who has performed a high number of procedures necessarily have good technical skills? We recognize the limitations of using simulated scenarios and a modified assessment tool to adapt to the virtual reality VATS lobectomy scenario, but modified assessment tools to better suit the research question have been used with very good results [7, 11, 34].

We aim to publish these first findings/experiences with the tool to show that initial validity evidence for a novel and unproven assessment tool designed for clinical use can be gathered using virtual reality simulation. Many studies describe the use of virtual reality simulators for assessing competence but most of these use simulator metrics as outcome parameters which makes it impossible to transfer the simulation-based assessment into real clinical practice. Our use of simulation for initial and standardized testing of a clinical assessment tool is quite novel and we would like to show that this tool—and probably many more already developed and tested assessment tools—can be used with accuracy on both simulated and real-life procedures. Furthermore, the novel assessment tool described in this study is designed to be used based on video recordings and we want to build on the literature showing that you can yield reliable results from video recordings, rendering the presence of a rater in the operating room unnecessary.

The VATSAT is designed to monitor progress of trainees' skills, facilitate structured feedback, contribute to measure the effect of training, and hopefully aid with certification in the future. The VATSAT is procedure-specific, can be used for blinded rating, and has high reliability which makes it useful with one rater. The ability to discriminate between novices and experienced surgeons was statistically significant, but to be used in the clinic where stakes are high for the trainee, the pass/fail score needs to be established for the full VATSAT and with ability to distinguish between experts and intermediates too. The next step will be to explore validity evidence for the original VATSAT's use in the operating room, as done in the just published paper by Petersen et al. [25].

Conclusion

Validity evidence was provided for a novel assessment tool for evaluating VATS lobectomy competence. The VAT-SAT provides supervisors and assessors with a procedure-specific assessment tool for evaluating VATS lobectomy performance and aids with the decision of when the trainee is ready for unsupervised performance.

Compliance with ethical standards

Disclosures Katrine Jensen's salary during her Ph.D. was partly funded by The Danish Cancer Society (Kræftens Bekæmpelse, "Knæk Cancer"), and this study was carried out as part of her Ph.D. Henrik Jessen Hansen and René Horsleben Petersen are at the Speakers Bureau of Medtronic. Kirsten Neckelmann, Henrik Vad, Lars Møller, Jesper Holst Pedersen, and Lars Konge have no conflicts of interest or financial ties to disclose.

References

1. Epstein RM, Cassel CK, Epstein RM, de Galan BE, van Gurp PJ, Stuyt PM (2007) Assessment in medical education. N Engl J Med 356:387–396. https://doi.org/10.1056/NEJMra054784
2. Lodge D, Grantcharov T (2011) Training and assessment of technical skills and competency in cardiac surgery. Eur J Cardio-Thorac Surg 39:287–293. https://doi.org/10.1016/j.ejcts.2010.06.035
3. Darzi A, Datta V, Mackay S (2001) The challenge of objective assessment of surgical skill. Am J Surg 181:484–486. https://doi.org/10.1016/S0002-9610(01)00624-9
4. Mcgaghie WC (2015) Mastery learning: it is time for medical education to join the 21st century. Acad Med 90:1–4. https://doi.org/10.1097/ACM.0000000000000911
5. Sharma B, Mishra A, Aggarwal R, Grantcharov TP (2011) Non-technical skills assessment in surgery. Surg Oncol 20:169–177. https://doi.org/10.1016/j.suronc.2010.10.001
6. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB (2011) Observational tools for assessment of procedural skills: a systematic review. Am J Surg 202:469–480. https://doi.org/10.1016/j.amjsurg.2010.10.020
7. Konge L, Lehnert P, Hansen HJ, Petersen RH, Ringsted C (2012) Reliable and valid assessment of performance in thoracoscopy. Surg Endosc 26:1624–1628. https://doi.org/10.1007/s00464-011-2081-7
8. Tong BC, Gustafson MR, Balderson SS, D'Amico TA, Meyerson SL (2012) Validation of a thoracoscopic lobectomy simulator. Eur J Cardio-Thorac Surg 42:364–369. https://doi.org/10.1093/ejcts/ezs012
9. Konge L, Larsen KR, Clementsen P, Arendrup H, Von Buchwald C, Ringsted C (2012) Reliable and valid assessment of clinical bronchoscopy performance. Respiration 83:53–60. https://doi.org/10.1159/000330061
10. Weizman NF, Manoucheri E, Vitonis AF, Hicks GJ, Einarsson JI, Cohen SL (2015) Design and validation of a novel assessment tool for laparoscopic suturing of the vaginal cuff during hysterectomy. J Surg Educ 72:212–219. https://doi.org/10.1016/j.jsurg.2014.08.015
11. Carlsen CG, Lindorff-Larsen K, Funch-Jensen P, Lund L, Charles P, Konge L (2014) Reliable and valid assessment of Lichtenstein hernia repair skills. Hernia 18:543–548. https://doi.org/10.1007/s10029-013-1196-2
12. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A (2008) Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. Ann Surg 247:372–379. https://doi.org/10.1097/SLA.0b013e318160b371
13. Yang SC, Merrill W (2014) Educational milestone development in phase II specialties: thoracic surgery. j Gr Med Educ 6:329–331
14. Messick S (1989) Meaning and values in test validation: the science and ethics of assessment. Educ Res 18:5–11. https://doi.org/10.3102/0013189x018002005
15. Reznick RK, MacRae H (2006) Teaching surgical skills–changes in the wind. N Engl J Med 355:2664–2669. https://doi.org/10.1016/S0084-392X(08)70199-0
16. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84:273–278. https://doi.org/10.1002/bjs.1800840237
17. Vogt VY, Givens VM, Keathley CA, Lipscomb GH, Summitt RL (2003) Is a resident's score on a videotaped objective structured assessment of technical skills affected by revealing the resident's identity? Am J Obstet Gynecol 189(3):688–691
18. Vassiliou MC, Feldman LS, Fraser S, Charlebois P, Chaudhury P, Stanbridge DD, Fried GM (2007) Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. Surg Innov 14:211–216. https://doi.org/10.1177/1553350607308466
19. Jensen K, Bjerrum F, Hansen HJ, Petersen RH, Pedersen JH, Konge L (2016) Using virtual reality simulation to assess competence in video-assisted thoracoscopic surgery (VATS) lobectomy. Surg Endosc. https://doi.org/10.1007/s00464-016-5254-6
20. Jensen K, Bjerrum F, Hansen HJ, Petersen RH, Pedersen JH, Konge L (2015) A new possibility in thoracoscopic virtual reality simulation training: development and testing of a novel virtual reality simulator for video-assisted thoracoscopic surgery lobectomy. Interact Cardiovasc Thorac Surg 21:420–426. https://doi.org/10.1093/icvts/ivv183
21. Konge L, Ringsted C, Bjerrum F, Tolsgaard MG, Bitsch M, Sørensen JL, Schroeder TV (2015) The simulation centre at Rigshospitalet, Copenhagen, Denmark. J Surg Educ 72:362–365. https://doi.org/10.1016/j.jsurg.2014.11.012
22. Jensen K, Petersen RH, Hansen HJ, Walker W, Pedersen JH, Konge L (2018) A novel assessment tool for evaluating competence in video-assisted thoracoscopic surgery lobectomy. Surg Endosc. https://doi.org/10.1007/s00464-018-6162-8
23. Subhi Y, Todsen T, Konge L (2014) An integrable, web-based solution for easy assessment of video-recorded performances. Adv Med Educ Pract 5:103–105. https://doi.org/10.2147/AMEP.S62277
24. Downing SM (2004) the metric of medical education reliability: on the reproducibility of assessment data. Med Educ 38:1006–1012. https://doi.org/10.1046/j.1365-2929.2004.01932.x
25. Petersen RH, Gjeraa K, Jensen K, Møller LB, Hansen HJ, Konge L (2018) Assessment of competence in Video Assisted Thoracoscopic Surgery (VATS) Lobectomy: a Danish nationwide study. J Thorac Cardiovasc Surg. https://doi.org/10.1016/j.jtcvs.2018.04.046
26. Crossley J, Davies H, Humphris G, Jolly B (2002) Generalisability: a key to unlock professional assessment. Med Educ 36:972–978. https://doi.org/10.1046/j.1365-2923.2002.01320.x
27. Konge L, Vilmann P, Clementsen P, Annema JT, Ringsted C (2012) Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal

staging of non-small cell lung cancer. Endoscopy 44:928–933. https://doi.org/10.1055/s-0032-1309892
28. Mahmood O, Dagnæs J, Bube S, Rohrsted M, Konge L (2017) Nonspecialist raters can provide reliable assessments of procedural skills. J Surg Educ. https://doi.org/10.1016/j.jsurg.2017.07.003
29. Våpenstad C, Buzink SN (2013) Procedural virtual reality simulation in minimally invasive surgery. Surg Endosc 27:364–377. https://doi.org/10.1007/s00464-012-2503-1
30. Walsh CM, Ling SC, Khanna N, Cooper MA, Grover SC, May G, Walters TD, Rabeneck L, Reznick R, Carnahan H (2014) Gastrointestinal endoscopy competency assessment tool: development of a procedure-specific assessment tool for colonoscopy. Gastrointest Endosc 79:798–807.e5. https://doi.org/10.1016/j.gie.2013.10.035
31. Gjeraa K, Spanager L, Konge L, Petersen RH, Ostergaard D (2016) Non-technical skills in minimally invasive surgery teams: a systematic review. Surg Endosc. https://doi.org/10.1007/s00464-016-4890-1
32. Hashimoto DA, Phitayakorn R, Fernandez-del Castillo C, Meireles O (2016) A blinded assessment of video quality in wearable technology for telementoring in open surgery: the Google Glass experience. Surg Endosc 30:372–378. https://doi.org/10.1007/s00464-015-4178-x
33. Vallurupalli S, Paydak H, Agarwal SK, Agrawal M, Assad-Kottner C (2013) Wearable technology to improve education and patient outcomes in a cardiology fellowship program—a feasibility study. Health Technol 3:267–270. https://doi.org/10.1007/s12553-013-0065-4
34. Hopmans CJ, den Hoed PT, van der Laan L, van der Harst E, van der Elst M, Mannaerts GHH, Dawson I, Timman R, Wijnhoven BPL, IJzermans JNM (2014) Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): a prospective multicenter study. Surgery 156:1078–1088. https://doi.org/10.1016/j.surg.2014.04.052

Affiliations

Katrine Jensen[1,2,6] · Henrik Jessen Hansen[1] · René Horsleben Petersen[1] · Kirsten Neckelmann[3] · Henrik Vad[4] · Lars Borgbjerg Møller[5] · Jesper Holst Pedersen[1] · Lars Konge[2]

1 Department of Cardiothoracic Surgery, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

2 Copenhagen Academy for Medical Education and Simulation (CAMES), University of Copenhagen and the Capital Region, Copenhagen, Denmark

3 Department of Cardiothoracic and Vascular Surgery, Odense University Hospital, Odense, Denmark

4 Department of Cardiothoracic and Vascular Surgery, Aarhus University Hospital, Skejby, Denmark

5 Department of Cardiothoracic Surgery, Aalborg University Hospital, Aalborg, Denmark

6 Department 5404, Copenhagen Academy for Medical Education and Simulation (CAMES), Blegdamsvej 9, 2100 Copenhagen, Denmark