View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by IUPUIScholarWorks

A video anchored rating scale leads to high inter-rater reliability of inexperienced and expert raters in the absence of rater training

Ronit Patnaik
Nicholas E. Anton
Dimitrios Stefanidis

Introduction

Laparoscopic intra-corporeal suturing (IS) is a crucial skill for many advanced laparoscopic procedures but mastery of this skill has a sharp learning curve. [1] This issue recently became quite apparent when fellowship directors raised concerns that many incoming surgery fellows were unprepared to perform IS adequately. [2] Accordingly, it is important to engage trainees in deliberate practice to refine their laparoscopic suturing skills, which has been found to contribute to enhanced suturing performance in the operating room. [3] Performance feedback is an important component of deliberate practice [4,5] and essential for surgical skill acquisition. To provide effective feedback, however, robust performance assessment is needed.

Global rating scales (GRS) and checklists are commonly used today for technical performance assessment.[6,7] Given the increasing interest nation-wide in using procedural video review for trainee performance assessment and feedback [8,9] the use of such assessment tools is only expected to increase. Nevertheless, such tools need to provide reliable assessments to promote robust skill acquisition. Unfortunately, there are a number of issues with their use. First, ratings are subject to multiple rater biases. For example, evaluators may perceive a resident's performance differently in light of their known PGY level or based on previous encounters either in or out of the operating room.[10] Importantly, the reliability and accuracy of assessments may be dependent upon rater characteristics, (e.g., familiarity with the scale, clinical expertise, training and personal idiosyncrasies) and on the complexity of the task, which lead some to question the use of expert global impressions in high-stakes assessment settings.[11] Indeed, evidence suggests that the IRR of GRSs and checklists tends to be variable and often poor. [12] A systematic review of the reliability of the objective structured clinical exam (OSCE) given to medical students by Brannick et al, found a large variability of IRR amongst raters. [13] To improve the IRR of assessment tools engaging raters with subject matter expertise and with rater training are recommended prior to their use. Using this approach, acceptable IRRs have been demonstrated previously. [6,14] Nevertheless, our group and others have found inadequate IRR even after rater training. [12,15] Further, while surgery faculty may be the ideal evaluators of surgical trainee performance, multiple demands on their time pose a significant barrier to their participation as raters and hampers adequate number of trainee performance assessments that is needed for effective feedback. [16,17] Allowing non-clinicians who have more time to devote to such assessments can, therefore, be beneficial. [18] Given that novice raters lack relevant clinical expertise, however, using them as raters presents a significant challenge and calls for innovative solutions.

Behaviorally anchored rating scales have been shown to be superior to plain number scales and ensure higher IRR by providing raters with written descriptions of each scale item. [19,20] Usually, visualization of a level of performance is intuitively more informative to a rater than a written description of the same performance. Our aim in this study was to evaluate the

---

This is the author's manuscript of the article published in final edited form as:

Patnaik, R., Anton, N. E., & Stefanidis, D. (2020). A video anchored rating scale leads to high inter-rater reliability of inexperienced and expert raters in the absence of rater training. American Journal of Surgery, 219(2), 221–226. https://doi.org/10.1016/j.amjsurg.2019.12.026

benefit of incorporating video anchors into an assessment tool for intracorporeal suturing we have developed [15] on inter-rater reliability of raters with different experience levels.

We hypothesized that incorporating video anchors to define performance into this tool would lead to high inter-rater reliability between novice, intermediate and expert raters. We also expected to find high correlation between these ratings and an objective IS score.

Materials and Methods

Performance Videos

After Institutional Review Board approval, 15 videos of laparoscopic IS performed on a live porcine Nissen fundoplication model (i.e., lasting a maximum of 10 minutes) were utilized in the present study. An expert in minimally invasive surgery (MIS) readied the porcine model for testing as described previously. [21-23] Performance videos only contained footage from the laparoscope and therefore included no images or audio of the operating room or the participants. Of note, the videos consisted of performances by novices, general surgery residents and an expert minimally invasive surgeon. The surgical novices were premedical students who had been trained to proficiency in laparoscopic suturing using the Fundamentals of Laparoscopic Surgery (FLS) IS training model but had never performed laparoscopic surgery. [24] The general surgery residents were all trained in laparoscopic suturing using the proficiency-based FLS-IS training curriculum and were represented by postgraduate years 1 to 5.

Participant performance on the 15 videos used in the study had been assessed previously using an objective suturing score.[21-23] This objective IS score was derived from the following equation: Performance score = Cutoff time (600 seconds) – Task completion time (seconds) – [10 x Accuracy error (millimeters outside black target on Penrose)] – [100 x Security error (secure knot = 0, partial slip = 1, knot failure = 2)]. An expert surgeon evaluated the accuracy and security errors at the time of IS performance. The 15 performance videos selected in this video had objective suturing scores in 5 different quintiles, from the lowest (0-120) to the highest (>480). Three videos from each quintile were utilized to ensure there were equal distributions of performance. All 15 videos were de-identified (i.e., assigned a randomly generated code) and randomized for the study.

Modified ISAT – Embedding short videos in the Intra-corporeal suturing assessment tool

The ISAT is a 23-item instrument that assesses IS performance via checklist items and 100-point visual analog scales for various domains related to IS as reported previously. The ISAT has displayed high correlation with the GOALS scale. The ISAT has the advantage of providing granular and global performance evaluations of trainee suturing ability utilizing performance videos. [15]

We modified the ISAT to include short video clips (<1 min) demonstrating clear examples of poor, average and excellent performance. An expert minimally invasive surgeon vetted these anchoring video clips after searching through our library of performance videos. For each of the eight criteria rated by a visual analog scale, video examples of poor, average and excellent performances were provided. A research assistant added text statements to the videos, which helped describe various suturing maneuvers and the intricacies of surgery. These text statements would allow even novices to better grasp the subtleties of suturing performance.

Some video examples of different performance levels (poor, average and excellent) that were used to define the rating scales can be found using the following links:

1) Poor needle handling: https://www.youtube.com/watch?v=iZNlIIlbKzc
2) Average needle handling: https://www.youtube.com/watch?v=nj11aYAEzco
3) Excellent needle handling: https://www.youtube.com/watch?v=z_DLVtJO0Qc

This modified ISAT was created on a REDCap survey to assist with data collection. [25] The modified ISAT was then used to assess 15 videos of trainees performing on a porcine fundoplication model. Excerpts from the original ISAT and modified ISAT can be found in Figure 1 and Figure 2 respectively.

Review process

The raters included 4 novices (medical or undergraduate students – Ages 20 – 25 years), 4 intermediates (residents, research fellows, and a performance coach- Ages 25 – 32 years) and 3 attending surgeons (Ages 35 – 45 years) with >10 years of experience. No rater training was provided prior to the study. The 15 performance videos were randomly assigned to the raters and they rated 5 videos every week, over a 3 week period. The flow of the study can be better visualized in Figure 3.

Blinding

To blind the laparoscopic videos such that the rater could not identity the performer, any identifiers were removed from the videos, and all 15 video clip files were labeled with a unique alphabet. To assign video reviews randomly, each reviewer received a unique, randomized list of video codes for each of the 3 weeks and was instructed to follow the order on their list when conducting reviews.

Statistical Analysis

To determine inter-rater reliability, the intraclass correlation and Pearson correlation coefficients were calculated among raters. We calculated the correlation amongst individuals and the intraclass correlation between different types of raters. We then compared the assessment scores of our groups of raters to the objective suturing assessment along with the Pearson correlation coefficient between each of the group of raters. For all analyses, a p-value of less than 0.05 was considered significant.

Results

All raters completed ratings for all 15 performance videos. The Pearson correlation coefficient amongst all individual raters was high at 0.80. The individual Pearson correlations are listed on table 1.

The intraclass correlation coefficient between ratings averaged over 11 raters was 0.98. The individual group ICC are listed in table 2.

Expert and intermediate ratings were highly correlated with the objective suturing scores, while this correlation was slightly lower for novice raters. In particular, this correlation was 0.84 for experts, 0.81 for intermediates, and 0.69 for novices.

Table 1: Pearson correlation coefficients between groups of raters

	R of ratings between groups
Novices vs Intermediates	0.74
Novice vs experts	0.75
Intermediates vs experts	0.82

Table 2: Intra-class correlation coefficient (ICC) within each group of raters

	ICC
Novices	0.84
Intermediates	0.96
Experts	0.94

Discussion

Multiple rater biases threaten the validity of assessment tools used for surgical performance assessments with increasing frequency today.[10] The literature reports significant variability among raters in their ratings even when using robust performance assessment tools.[13] While rater expertise and training can decrease this variability and improve IRR, these approaches are associated with a number of challenges: engagement of expert surgeons in conducting performance reviews and assessments and offering them rater training is time consuming and costly. Further, even if they commit, expert surgeons tend to be late in completing their ratings [18] preventing the timely provision of feedback to trainees. The challenges with using faculty raters have fueled the development of a new industry of crowd-sourced raters.[26,27] Improving the reliability of assessment tools even in the absence of rater training so that they can be also used by non-faculty raters would be highly desirable.

To address these issues, we incorporated video-based behavioral anchors in a intracorporeal suturing performance assessment tool and studied their impact on the IRR of raters with varying surgical expertise. We found that the correlation amongst individual raters was high at 0.80 despite their diverse expertise. Importantly, the interclass correlation coefficient (ICC) between ratings provided by the 11 raters of this study was 0.98. The high correlation of the provided ratings on the ISAT with the objective performance scores provides further support for the value of the video anchors. And despite the expert ratings having the highest correlation with the objective suturing scores, ratings of surgery residents had very similar correlations and even novice raters came close. These findings are in sharp contrast with our prior experience using the ISAT tool without video anchors. In a previous study conducted by our group[16], we found that the IRR between novice and expert raters using the same ISAT tool but without the videos was markedly lower than in the current study (ICC of 0.68). For expert raters the ICC was 0.69 and for novices it was 0.71. We did not include intermediate raters in that study. These findings are promising as they suggest that the addition of videos leads to improved ICC among raters but given that this comparison is indirect, a head to head comparison is needed to demonstrate the value of the addition of video anchors to the original ISAT.

Low inter-rater reliabilities when using performance assessment tools has been reported previously. [12] In a 2019 study by Satava et al about the fundamentals of robotic surgery skills curriculum, the GEARS (Global Evaluative Assessment of Robotic Skills) tool had a low IRR (0.40 – 0.67) despite prior rater training. [28] Cooper et al developed a tool to assess teamwork in

an emergency resuscitation; in their study, they found a low inter-rater reliability of 0.55. [29] Thus, despite not offering rater training in our study we were able to achieve higher IRR than prior studies that used rater training.

The literature supports our approach of using non-medically trained individuals to assess surgical skills. [18,27]. Mahmood et al showed that trained non-specialist raters can provide reliable and valid assessments of video recorded cystoscopies compared to specialist raters as evidenced by similar internal consistency, inter-rater reliability, and test-retest reliability. [30] Yeung et al showcased that raters with varying levels of expertise can reliably grade performance of an intra-corporeal suturing task. While novices were less confident in their grading, both groups were able to assign comparable scores and identify similar elements of suturing skill as being important in terms of assessment.[31] Similar to us, they did not provide rater training. In their study, raters needed to watch on average, 13.8 videos before they could confidently grade, however. One of the main rationales for adding videos to our ISAT tool was to see if the need for pre-training the raters could be avoided in addition to improving the inter-rater reliability. The modified ISAT provides visual cues and explanations for what raters should look for while evaluating videos to improve the accuracy of their ratings. Our study shows that, with adequate guidance through video-based anchors of performance, novices may be able to provide reliable performance ratings similar to more experienced raters even in the absence of training. Enlisting novice raters utilizing the modified ISAT may reduce the burden on expert surgeons, who are already stretched thin with several commitments.

There are some limitations in our work that are worth highlighting and will likely affect future work in this area. We did not compare ratings on the modified ISAT with the video anchors to ratings on the original ISAT without video anchors directly. Further work in this area is needed and is planned. In addition, incorporating video anchors to other performance assessment tools besides ISAT may demonstrate whether effect observed in this study is generalizable. We should also note that the video enhanced ISAT tool may be most effectively used during asynchronous video-based reviews as real-time assessments may be challenging. Further, the definition of expert, intermediate and novice evaluators was based on clinical expertise and may not accurately reflect expertise differences in rating ability. Nevertheless, all our expert raters were attending surgeons who assess resident performance on a regular basis (in our institution all faculty regularly complete evaluations on residents using the System for Improving and Measuring Procedural Learning (SIMPL) tool[32] among others) and most of them also have had prior experience with video assessment. Some of our intermediates also had previous experience with performance rating using video while no novice had prior rating experience. We, therefore, believe that comparatively, our attendings could also qualify as rating experts in comparison with the intermediate and novice raters and the intermediate raters had more experience than the novices. Thus, we believe that the rater assignments into the stated groups was valid. Finally, we had a low number of participating raters in each group – 4 novices, 4 intermediates and 3 experts. Despite this we were able to show high IRR among all raters and rater groups.

Conclusions

Incorporating video anchors to define performance into the intra-corporeal suturing assessment tool (ISAT) led to high inter-rater reliability and enabled less experienced raters without prior rater training to achieve similar consistency in their ratings as experts. These

findings support the incorporation of video based anchors in performance assessment tools as an effective method to ensure reliable ratings independent of rater experience.

References

1. Stefanidis D, Hope WW, Korndorffer Jr JR, Markley S, Scott DJ. Initial laparoscopic basic skills training shortens the learning curve of laparoscopic suturing and is cost-effective. *Journal of the American College of Surgeons.* 2010;210(4):436-440.

2. Mattar SG, Alseidi AA, Jones DB, et al. General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. *Annals of surgery.* 2013;258(3):440-449.

3. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. *Acad Emerg Med.* 2008;15(11):988-994.

4. Hattie J, Timperley H. The power of feedback. *Review of educational research.* 2007;77(1):81-112.

5. Burlew CC. Surgical education: Lessons from parenthood. *The American Journal of Surgery.* 2017;214(6):983-992.

6. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation‐based assessment. *Medical education.* 2015;49(2):161-173.

7. Norman G, Van der Vleuten C, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical education.* 1991;25(2):119-126.

8. Soucisse ML, Boulva K, Sideris L, Drolet P, Morin M, Dubé P. Video coaching as an efficient teaching method for surgical residents—a randomized controlled trial. *Journal of surgical education.* 2017;74(2):365-371.

9. Herrera-Almario GE, Kirk K, Guerrero VT, Jeong K, Kim S, Hamad GG. The effect of video review of resident laparoscopic surgical skills measured by self-and external assessment. *The American Journal of Surgery.* 2016;211(2):315-320.

10. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. *The American journal of surgery.* 2007;193(5):551-555.

11. Lievens F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology.* 2001;86(2):255.

12. Mazor KM, Ockene JK, Rogers HJ, Carlin MM, Quirk ME. The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Advances in health sciences education.* 2005;10(1):37-51.

13. Brannick MT, Erol‐Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Medical education.* 2011;45(12):1181-1189.

14. Norman G. Editorial–checklists vs. ratings, the illusion of objectivity, the demise of skills and the debasement of evidence. *Advances in Health Sciences Education.* 2005;10(1):1-3.

15. Anton NE, Sawyer JM, Korndorffer Jr JR, et al. Developing a robust suturing assessment: validity evidence for the intracorporeal suturing assessment tool. *Surgery.* 2018;163(3):560-564.

16. Sawyer JM, Anton NE, Korndorffer Jr JR, DuCoin CG, Stefanidis D. Time crunch: increasing the efficiency of assessment of technical surgical skill via brief video clips. *Surgery.* 2018;163(4):933-937.

17. Acton RD, Chipman JG, Lunden M, Schmitz CC. Unanticipated teaching demands rise with simulation training: strategies for managing faculty workload. *Journal of surgical education.* 2015;72(3):522-529.

18. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. *JAMA surgery.* 2015;150(11):1086-1087.

19. Devcich DA, Weller J, Mitchell SJ, et al. A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: development and initial evaluation of the WHOBARS. *BMJ Qual Saf.* 2016;25(10):778-786.

20. Ohland MW, Layton RA, Loughry ML, Yuhasz AG. Effects of behavioral anchors on peer evaluation reliability. *Journal of Engineering Education.* 2005;94(3):319-326.

21. Prabhu A, Smith W, Yurko Y, Acker C, Stefanidis D. Increased stress levels may explain the incomplete transfer of simulator-acquired skill to the operating room. *Surgery.* 2010;147(5):640-645.

22. Stefanidis D, Korndorffer JR, Markley S, Sierra R, Heniford BT, Scott DJ. Closing the gap in operative performance between novices and experts: does harder mean better for laparoscopic simulator training? *Journal of the American College of Surgeons.* 2007;205(2):307-313.

23. Korndorffer JR, Dunne JB, Sierra R, Stefanidis D, Touchard CL, Scott DJ. Simulator training for laparoscopic suturing using performance goals translates to the operating room. *Journal of the American College of Surgeons.* 2005;201(1):23-29.

24. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *The American journal of surgery.* 2010;199(1):115-120.

25. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics.* 2009;42(2):377-381.

26. Holst D, Kowalewski TM, White LW, et al. Crowd-sourced assessment of technical skills: an adjunct to urology resident surgical simulation training. *Journal of endourology.* 2015;29(5):604-609.

27. Chen C, White L, Kowalewski T, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research.* 2014;187(1):65-71.

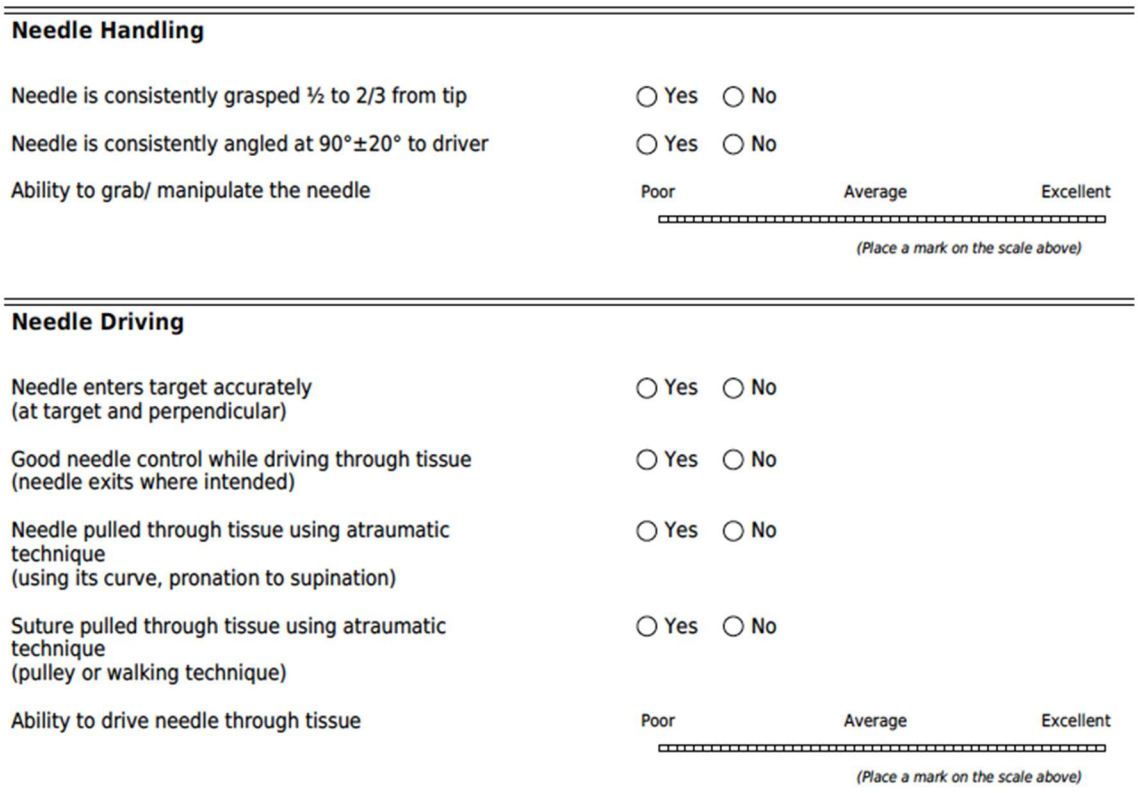
28. Satava RM, Stefanidis D, Levy JS, et al. Proving the Effectiveness of the Fundamentals of Robotic Surgery (FRS) Skills Curriculum: A Single-blinded, Multispecialty, Multi-institutional Randomized Control Trial. *Annals of surgery.* 2019.

29. Cooper S, Cant R, Porter J, et al. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation.* 2010;81(4):446-452.

30. Mahmood O, Dagnæs J, Bube S, Rohrsted M, Konge L. Nonspecialist raters can provide reliable assessments of procedural skills. *Journal of surgical education.* 2018;75(2):370-376.

31. Yeung C, Carrillo B, Pope V, Hosseinpour S, Gerstle JT, Azzie G. Video assessment of laparoscopic skills by novices and experts: implications for surgical education. *Surgical endoscopy.* 2017;31(10):3883-3889.

32. Gunderson K, Sullivan S, Warner-Hillard C, et al. Examining the Impact of Using the SIMPL Application on Feedback in Surgical Education. *Journal of surgical education.* 2018;75(6):e246-e254.



Needle Handling

Needle is consistently grasped ½ to 2/3 from tip	◯ Yes	◯ No	
Needle is consistently angled at 90°±20° to driver	◯ Yes	◯ No	
Ability to grab/ manipulate the needle	Poor	Average	Excellent

(Place a mark on the scale above)

Needle Driving

Needle enters target accurately (at target and perpendicular)	◯ Yes	◯ No
Good needle control while driving through tissue (needle exits where intended)	◯ Yes	◯ No
Needle pulled through tissue using atraumatic technique (using its curve, pronation to supination)	◯ Yes	◯ No
Suture pulled through tissue using atraumatic technique (pulley or walking technique)	◯ Yes	◯ No
Ability to drive needle through tissue	Poor	Average Excellent

(Place a mark on the scale above)

Figure 1: Excerpt from original ISAT

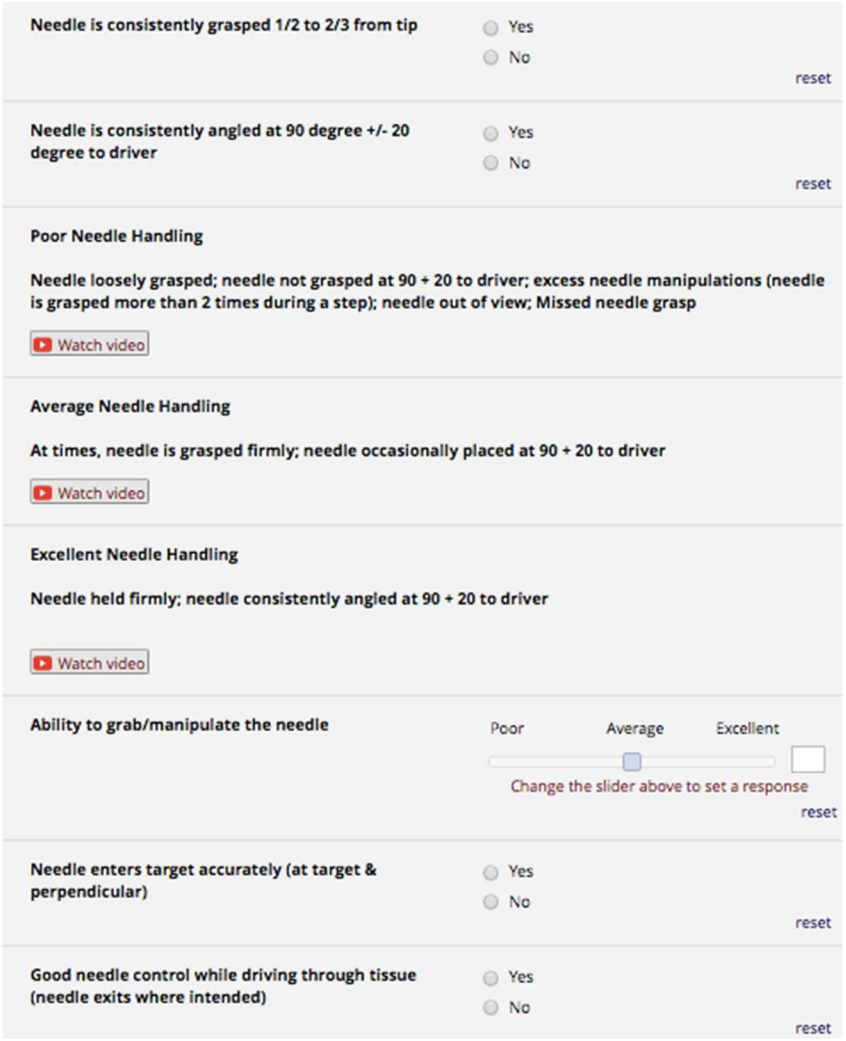


Figure 2: Excerpt from modified ISAT

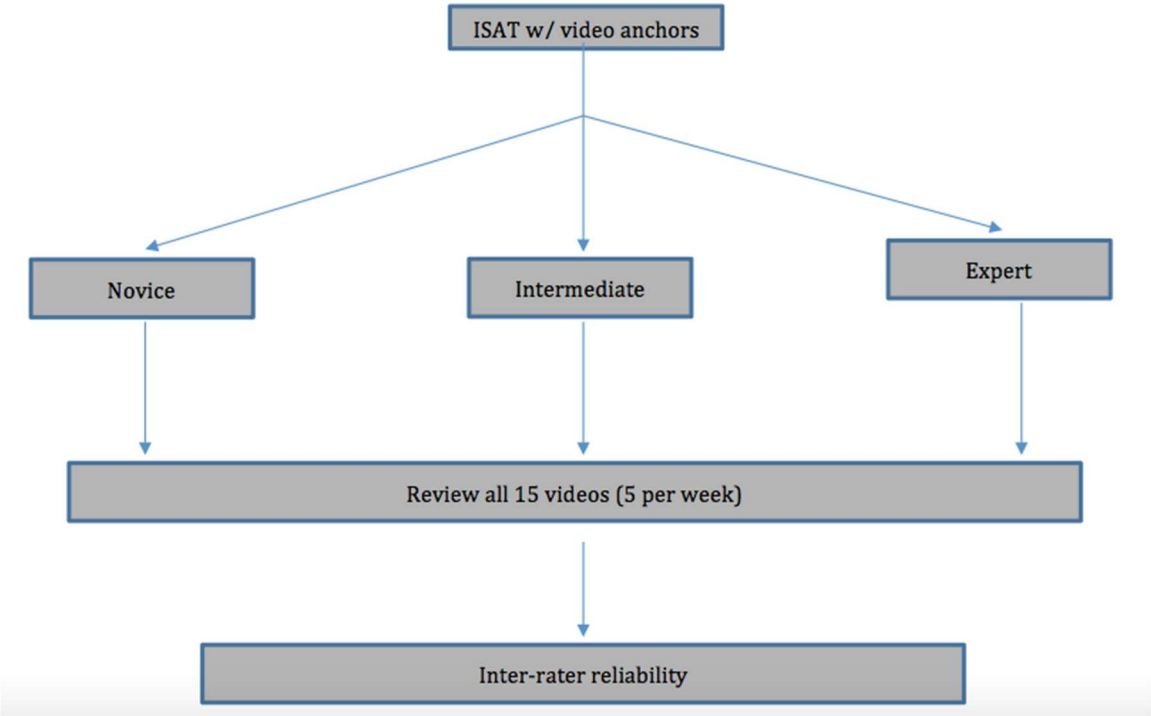


Figure 3: Flow of the study

Needle Handling

Needle is consistently grasped ½ to 2/3 from tip	◯ Yes ◯ No
Needle is consistently angled at 90°±20° to driver	◯ Yes ◯ No
Ability to grab/ manipulate the needle	Poor — Average — Excellent *(Place a mark on the scale above)*

Needle Driving

Needle enters target accurately (at target and perpendicular)	◯ Yes ◯ No
Good needle control while driving through tissue (needle exits where intended)	◯ Yes ◯ No
Needle pulled through tissue using atraumatic technique (using its curve, pronation to supination)	◯ Yes ◯ No
Suture pulled through tissue using atraumatic technique (pulley or walking technique)	◯ Yes ◯ No
Ability to drive needle through tissue	Poor — Average — Excellent *(Place a mark on the scale above)*

Figure 1: Excerpt from original ISAT

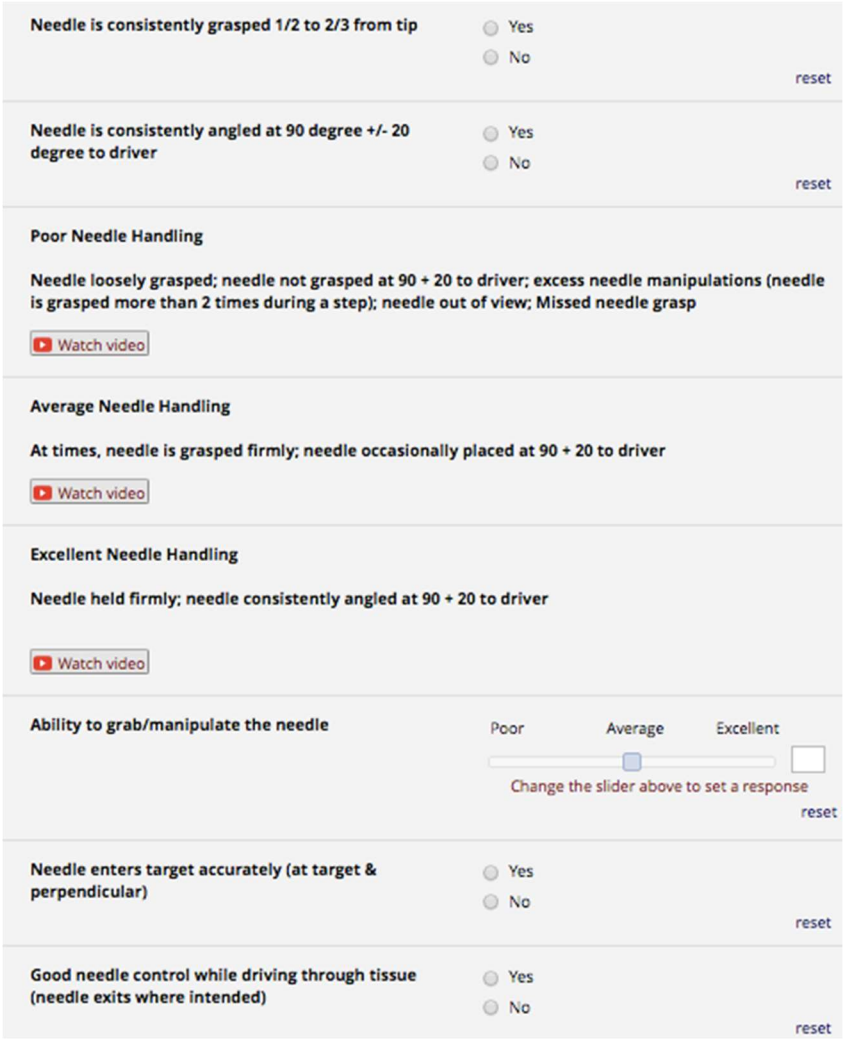


Figure 2: Excerpt from modified ISAT

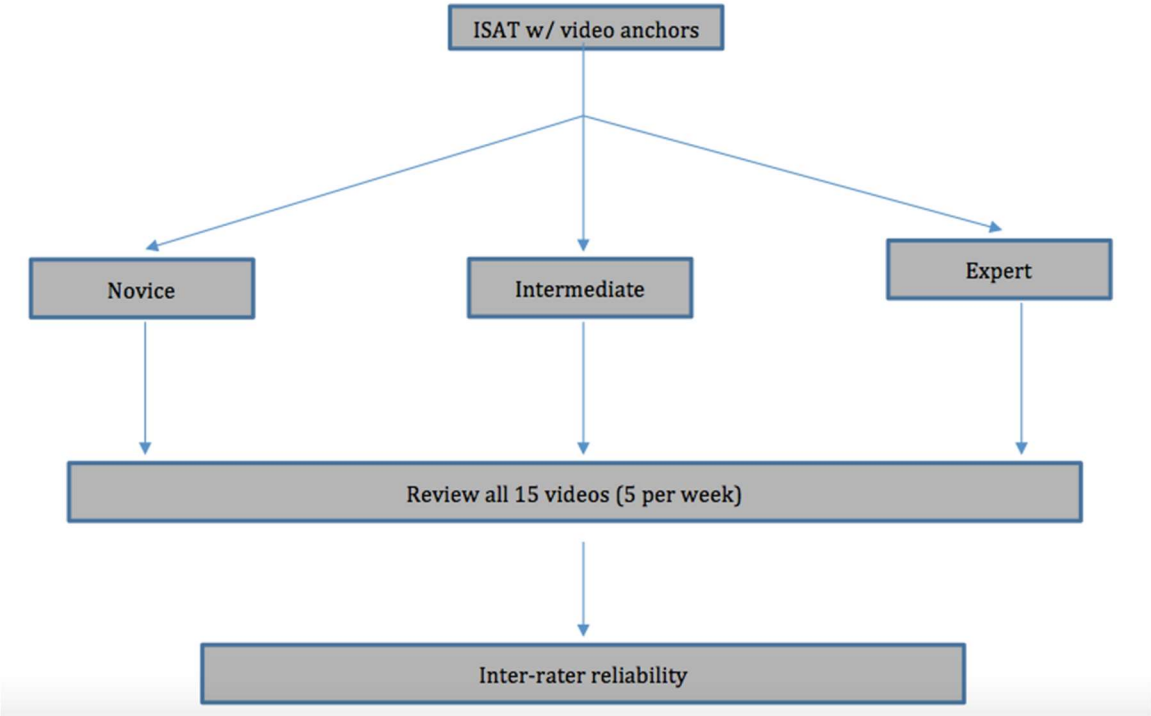


Figure 3: Flow of the study