

Video Observed Surgical Skills vs Directly Observed Surgical Skills Among Postgraduate Residents of Pediatric Surgery: A Comparative Study

Muhammad Sharif

King Edward Medical University

Muhammad Adil Iftikhar

King Edward Medical University

Batool Fatima

dr.batoolfatima@gmail.com

King Edward Medical University <https://orcid.org/0009-0000-3639-7183>

Masood Jawaid

The University of Lahore

Muhammad Soban Sharif

King Edward Medical University

Jaweria Sharif

Allama Iqbal Medical College

Research Article

Keywords: Surgical Procedures, Operative, Internship and Residency, Clinical Competence, Educational Measurement, Video Recording, Pediatric Surgery

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11234023/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Purpose

Pediatric surgery residents must attain advanced surgical skills, which requires constant supervision and assessment. This study compared video observed surgical skills (VOSS) with directly observed surgical skills (DOSS) among postgraduate pediatric surgery residents.

Methods

A prospective comparative study was conducted at the Pediatric Surgery Department, King Edward Medical University. Twenty final-year residents performed 96 elective procedures. Each was observed directly by two faculty members (DOSS) and video recorded; two other faculty members later scored the recordings (VOSS) using the same OSATS proforma.

Results

Mean total OSATS score was 24.99 (SD 2.46) for DOSS and 20.89 (SD 4.98) for VOSS, a mean difference of -4.11 points ($p < 0.001$; Wilcoxon signed rank $p < 0.001$). In contrast, all seven individual OSATS domains met the prespecified equivalence margin of ± 1 point, with mean differences from -0.15 (use of assistants) to -0.33 (respect for tissue) and 90% confidence intervals within ± 1 for every domain (TOST $p < 0.001$).

Conclusion

VOSS showed equivalence with DOSS across all seven OSATS domains among final-year pediatric surgery residents, supporting its use as a practical alternative for domain-level formative assessment. Video assessors scored residents lower overall, so VOSS should be used with caution in high-stakes or borderline competency decisions until procedure-specific tools and further validation are available. By removing the need for faculty presence in theatre, VOSS may make workplace-based assessment more feasible in resource-limited programs.

Introduction

Pediatric surgery is a highly specialized field that requires extensive training and skill development. Postgraduate residents in pediatric surgery are supposed to gain advanced surgical skills to perform complex procedures in a safe and effective manner that needs constant supervision. For the acquisition of fine surgical skills for a specific surgical procedure trainee have to perform surgeries at different levels to attain a competency. [1] Final stage of acquisition of surgical skills is to perform surgeries

independently. [2] At this level trainee should be observed via an assessment tool and mechanism so that effective feedback can be provided for further improvement. [3]

The mechanism of assessment can be direct observation or via video recordings. The Directly Observed Surgical Skills (DOSS) technique is adopted by institutes as a valid method of formative assessment of postgraduate trainees. Whereas assessment using video recordings is gaining popularity because DOSS evaluation will not come in handy in terms of times and available human resources.[4] This needs the faculty presence in the operation theaters and also multiple trainees can not be assessed at one time. A study comparing video observed procedural skills (VOPS) and directly observed procedural skills (DOPS) for ophthalmology residents shows a strong positive correlation between the total scores of the "task-specific" ($r = 0.64$, $P < 0.001$) and "global" ($r = 0.38$, $P = 0.003$) indices when comparing VOPS with DOPS. [5] Narges and her colleagues conducted a study with the objective of comparing the outcomes of the VOSS technique with the widely employed Direct Observation of Procedural Skills (DOSS) method for ophthalmology residents performing trabeculectomies. The Bland-Altman analysis revealed that all data points were within the 95% margins of agreement, which ranged from -1.46 to 0.46. The study concludes that VOSS is a viable, valid, and dependable evaluation approach. [6]

Various grading scales have been devised to enhance the structure and objectivity of the assessment process for surgical technical competency. Several rating scales are employed to assess surgical technical skills. The Global Rating Scales (GRS), the 'Objective Structured Assessment of Technical Skills' (OSATS), and the 'Global Evaluative Assessment of Robotic Skills' (GEARS) are used for a general evaluation of surgical technical skills. [7] The tools used for assessment are multiple but for surgical trainees OSAT is widely used. Scores are calculated from seven categories generic OSAT (objective structured assessment of technical skills) global rating scale on a Likert scale from 1-5 out of total 35 numbers. [8,9]

Primarily this study is expected to show effectiveness of VOSS assessment method as alternative to DOSS assessment method. Assessment by Video observation is a potential alternative to assessment by direct observation in surgical skills training. Assessment time by faculty will be reduced, surgical skill can be assessed in more critical way by repeated observations in video and improved surgical training programs can be structured through frequent feedback. [10]

Methods

The study was conducted at the pediatric surgery department of King Edward Medical University/Mayo Hospital. Ethical approval was taken from the IERB department of King Edward Medical University # 190/RC/KEMU dated 8th June 2023. Duration of the study was August 2023 till July 2024.

The sample size was calculated using a one-sided significance level of 5%, power of 80%, an expected mean difference of zero, and equivalence limits of -1.5 to $+1.5$ points on the total 35-point OSATS score,

36 completed paired resident-procedure assessments were required using an exact t-distribution-based paired equivalence calculation. The standard deviation of the paired differences was estimated from the study by Ghiasian et al. in which directly observed and video-observed global surgical performance scores had standard deviations of 2.40 and 2.19, respectively, with a correlation coefficient of 0.38. Paired difference standard deviation was 2.99 for 35 points OSATS. [5] The minimum sample size calculated using the formula from a study by Wang et al. is calculated to be 34.

The participants included residents who performed elective pediatric surgical procedures such as circumcision, surgery to repair an undescended testis, and repair of an inguinal hernia. Only residents in their fourth and last year of pediatric surgery training were eligible to participate. Surgeries which resulted in complications leading to prolonged duration (more than 45 min) were excluded and emergency surgeries done during elective morning list were also excluded. After written informed consent two faculty members watched surgical procedures first hand and concurrently recorded them on film. Two more faculty members later evaluated the recorded films, after removing the identifiers of the participants, using the same evaluation criteria as the in-person observation.

Data was collected on a detailed DOSS assessment form of Generic OSATS global rating scale. It has a likert scale scoring system for 7 criterias that are judged during surgery, which include respect for tissue, time & motion, instrument handling, knowledge of instruments, use of assistants, and flow of operation and forward planning, table 1.

	1	2	3	4	5
Respect for tissue	Frequently used unnecessary force on tissue or caused Damage by inappropriate use of instruments	Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately with minimal damage	
Time and motion	Many unnecessary moves	Efficient time/motion but some unnecessary moves		Economy of movement and Maximum efficiency	
Instrument handling	Repeatedly makes tentative or awkward moves with instruments	Competent use of instruments although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness	
Knowledge of instruments	Frequently asked for the wrong instrument or used an inappropriate instrument	Knew the names of most instruments and used appropriate instrument or the task		Obviously familiar with the instruments required and their names	
Use of assistants	Consistently placed assistants poorly or failed to use assistants	Good use of assistants most of the time		Strategically used assistant to the best advantage at all Times	
Flow of operation and forward planning	Frequently stopped operating or needed to discuss next move	Demonstrated ability for forward planning with steady progression of operative procedure		Obviously planned course of operation with effortless flow from one move to the next	

Knowledge of specific procedure	Deficient knowledge. Needed specific instruction at most operative steps	Knew all important aspects of the operation	Demonstrated familiarity with all aspects of the operation
--	--	---	--

Table-1. Generic OSATS Global Rating Scale

Normality of the OSATS scores by direct and video assessors were checked by Shapiro Wilk test. Normally distributed data was presented as means with SD and skewed data shall be presented asmedian with IQR. To check the agreemenet between VOSS and DOSS, Two One Sided Tests were applied to check if the mean difference between the scores fall between 90% confidence interval. Interclass Correlation Coefficient >0.75 indicated excellent grading alignment. Bland-Altman scatter plot was used to visually demonstrate the scoring variations.

Results

Total 96 procedure were assessed after excluding missing data values and technical issues in video recordings. Two assessor scores for DOSS and 2 assesor scores for VOSS were compared as paired procedures. Shapiro-Wilk testing indicated that the paired differences were not normally distributed for the total score or for any domain (all p<0.01), likely reflecting the discrete, ceiling affected nature of the 1 to 5 OSATS scale.

Mean total OSATS score was 24.99 (SD 2.46) for DOSS and 20.89 (SD 4.98) for VOSS, a mean difference of -4.11 points (VOSS lower than DOSS, paired t test p<0.001; Wilcoxon signed rank p<0.001). The 90% confidence interval for this difference, -5.18 to -3.04, fell entirely outside the prespecified total score equivalence margin of ±3 points, so equivalence could not be demonstrated by two one sided test (TOST) analysis (p=0.955). (Table 2, figure 1)

Table 2. Descriptive comparison of DOSS and VOSS scores (n = 96)

OSATS domain	DOSS mean (SD)	VOSS mean (SD)	Mean difference (VOSS – DOSS)	Paired t test p value
Total score	24.99 (2.46)	20.89 (4.98)	–4.11	<0.001
Respect for tissue	3.18 (0.90)	2.85 (0.88)	–0.33	0.008
Time and motion	3.53 (0.72)	3.32 (0.83)	–0.21	0.021
Instrument handling	3.55 (0.64)	3.30 (0.84)	–0.25	0.008
Knowledge of instruments	3.60 (0.61)	3.41 (0.82)	–0.20	0.028
Use of assistants	3.51 (0.73)	3.36 (0.83)	–0.15	0.125
Flow and forward planning	3.64 (0.53)	3.35 (0.78)	–0.28	0.001
Knowledge of procedure	3.61 (0.64)	3.36 (0.88)	–0.25	0.024

SD, standard deviation. DOSS, directly observed surgical skills. VOSS, video observed surgical skills.

In contrast to the total score, all seven individual OSATS domains met their prespecified equivalence margin of ± 1 point. Mean differences ranged from –0.15 (use of assistants) to –0.33 (respect for tissue), and the 90% confidence interval of the mean difference was fully contained within ± 1 for every domain (all TOST $p < 0.001$). Summed across domains, these small unidirectional differences accumulated into a total-score gap that exceeded the wider margin set for the composite score (figure 2).

Limits of agreement for the domains were correspondingly narrow, for example –2.64 to 1.99 for respect for tissue and –1.90 to 1.34 for flow and forward planning (Table 3). A small but statistically significant proportional bias was present for instrument handling, knowledge of instruments, flow and forward planning, and knowledge of procedure (all $p \leq 0.007$), while respect for tissue, time and motion, and use of assistants showed no evidence of proportional bias (all $p > 0.15$).

Table 3. Bland-Altman agreement between DOSS and VOSS scores

OSATS domain	Bias (95% CI)	Lower LoA	Upper LoA	Proportional bias p value
Total score	-4.11 (-5.39 to -2.83)	-16.52	8.30	<0.001
Respect for tissue	-0.33 (-0.57 to -0.09)	-2.64	1.99	0.815
Time and motion	-0.21 (-0.39 to -0.03)	-1.95	1.53	0.157
Instrument handling	-0.25 (-0.43 to -0.07)	-2.03	1.53	0.007
Knowledge of instruments	-0.20 (-0.37 to -0.02)	-1.90	1.50	0.002
Use of assistants	-0.15 (-0.33 to 0.04)	-1.96	1.66	0.212
Flow and forward planning	-0.28 (-0.45 to -0.11)	-1.90	1.34	<0.001
Knowledge of procedure	-0.25 (-0.46 to -0.03)	-2.29	1.80	0.002

LoA, 95% limits of agreement (bias $\pm$ 1.96 SD of the differences). Proportional bias p value is from the Pearson correlation between the DOSS-VOSS mean and the difference.

Bland-Altman analysis showed the same negative bias, with wide 95% limits of agreement (-16.52 to 8.30) and a significant proportional bias ($r=0.64$, $p<0.001$): the VOSS-DOSS difference became less negative as the pair mean score increased, meaning video assessment scored the weaker performers relatively lower than direct assessment did, while agreement was closer among the stronger performers (Figure 3). This suggests that video-based assessment may be less generous, or direct assessment more generous, specifically when rating residents who are struggling, which has implications for using VOSS in high-stakes or borderline competency decisions.

Discussion

The DOSS is a conventional assessment method that depends on the examiner, evaluating the trainee in real-time and offering fast, instant feedback as well as beneficial educational consequences. [12,13] Live assessment in DOSS also makes it possible to take into account every aspect of the procedures, including the surgeon's charge, the operating team's verbal and nonverbal communication, and equipment errors that occur throughout the procedures. [12] There have been certain DOSS documented flaws. With the DOSS approach, assessors are not able to conceal the identity of the trainee, which increases the possibility of observer bias. Moreover, trainees may experience tension and anxiety due to the assessor's time-consuming stay in the operating room, which could negatively impact their typical performance in DOSS.

For workplace-based assessment, VOSS was later established and offers several advantages over DOSS. The fact that VOSS reduces the time-consuming nature and coordination of the assessor's schedules to be present in the operating room is one of its most significant advantages. Raters are free to evaluate whenever they have enough time. Thus, the VOSS approach reduces rater burnout from exhaustion and lack of focus in addition to increasing efficiency. Sufficient time and appropriate conditions for an accurate, step-by-step evaluation of the process enable more thorough identification of any mistakes in VOSS, which results in lower scores as compared to DOSS. [6] Additionally, this approach provides anonymity, which removes observational bias and improves evaluation objectivity, it reduces the nervousness of the trainees during the assessment. [14] Video based assessment is being considered as part of continuous certification program for general surgeons and their relevant specialities by The American Board of Surgery. [15]

Supervisors typically provide a single global rating for surgical examinations. [16] Although they cannot be regarded as a perfect evaluation of technical skill, global scores are more broadly applicable and show superior reliability and validity because they rate the residents' skills and the adequacy of their technical proficiency using general performance as a standard. On the other hand, procedure-specific task-specific scores necessitate watching the full surgery. By recognizing the shortcomings of the trainees, they give them more formative criticism. [17]

We used 5 point generic global rating scale for assessment. While in another study 6 item global rating assessment tool was used. [5] Results of some other studies conclude that video recorded assessment is better than directly observational assessment. [18,19,20]

One advantage of video observed surgical assessment is that recorded video can be used for feedback and improvement as described in one study. These videos can later be assessed by peers and the outcome of feedback is studied in many studies. These recordings can also be evaluated by trained AI models to assess and feedback, which is a leap in surgical coaching. (20) According to another study Video observed surgical skill assessed by expert faculty members improved future surgical skills by giving feedback to trainees about their mistakes and score. [19]

Feedback is also taken from the trainees about their assessment as done by a pilot study in India for surgical trainees on their workplace based assessment. [21] In the next phase of our study the feedback from trainees can be taken about video observation and direct observation both to compare the results. This will surely give meaningful insight to assess the comfort and understanding of trainees. [22]

This study shows that assessing pediatric surgical trainees from recorded video (VOSS) gives essentially the same result as assessing them in the operating room in real time (DOSS). Total and domain-level OSATS scores fell within the prespecified equivalence margins on two one-sided testing, and Bland–Altman analysis showed acceptable agreement between the two methods with no systematic bias. Beyond the practical gain of freeing assessment from faculty availability and theatre scheduling, this equivalence implies that the operative video carries the information an expert relies on to judge performance. That information, in principle, is also available to a machine. [20]

This is the premise on which automated surgical skill assessment has been built. Yanik et al. trained a deep learning model on raw video of laparoscopic suturing trials and were able to classify pass–fail outcomes with an accuracy of roughly 0.88, and to predict performance scores with a Spearman correlation of about 0.75, without any human rater involved. [23] Liu et al. went further, proposing a framework that assesses skill by combining several different signals from the operative video – how instruments are used, what intraoperative events occur, and other indirect markers of competence – and showed that drawing on multiple aspects together performs better than relying on any one alone. This will ultimately lead to the assessment by AI, of how well the surgery was performed. [24]

Benmansour et al. addressed a different limitation, namely that most automated systems sort surgeons into a few discrete categories and cannot express the gradations in between. Their model instead produced continuous OSATS scores on a 1-to-5 scale, achieving Spearman correlations of 0.82, 0.60 and 0.65 for knot-tying, needle-passing and suturing. It did so, however, from recorded instrument motion rather than from video, a reminder that the field's more refined scoring methods still tend to depend on platforms that track instruments directly. [25]

Taken together, these studies describe a field that is technically capable but narrowly trained. Almost all of this work rests on laparoscopic, robotic or simulator data, for the straightforward reason that these platforms produce clean, consistent recordings from a fixed camera, or record instrument motion directly. Open surgery is barely represented. Yet open surgery is what most of the world's operations still are, and it is how residents in low- and middle-income countries are trained, examined and certified. The trainee who most needs scalable, objective feedback is the one the existing models cannot assess. [26]

Our findings suggest a way around this. If video alone is enough to reproduce a valid assessment – as the equivalence between VOSS and DOSS demonstrates – then the recording does not need to come from a modular theatre with installed cameras. A handheld or mobile-phone recording of an open case may serve just as well, both as material to be assessed and as data on which to train. Scores accumulated across cases would produce an individual learning curve, letting programmes see how each trainee is progressing over time and identify those who need closer supervision. Where neither theatre camera systems nor spare faculty assessor time exist, the ability to obtain valid assessment from a phone recording of an open operation offers a realistic path toward competency-based surgical training.

Strengths of study: Strengths of our study is that this is first study about video observed surgical skill regarding paediatric surgery in Pakistan. We used Generic OSAT global rating scale which has reliability and validity of 0.61-0.74 and 3.55 to 4.45 respectively. Relatively larger sample size is assessed as compared to other studies. We have involved four faculty members as assessor to reduce bias element by taking mean score of assessors both in DOSS and VOSS methods. Also, such assessment may be added in portfolio replacing operative log as better assessment tool.

Limitations of study: There are some limitations in our study. This is weak experimental study as we did no randomisation. Quality of camera may affect assessment level by faculty members, There was

difficulty in finding and transferring video to Lap top and WhatsApp if we use high definition camera. We also do not have facility of in built video camera in operation light which might have recorded video in a better way. We used generic global rating scale for assessment but it will be better if we create checklist and procedure specific rating scale for paediatric surgery which will result in better outcome of assessment in paediatric surgery examinations. We also did not repeat assessment by recorded video after giving feedback to students about their deficiencies in performance during surgery. So there is need to do further studies using checklist and procedure specific rating scale especially meant for paediatric surgery

References

1. Bjerrum F, Thomsen ASS, Nayahangan LJ, Konge L. Surgical simulation: current practices and future perspectives for technical skills training. *Med Teach*. 2018;40(7):668-75.
<https://doi.org/10.1080/0142159X.2018.1472754>
2. Steiman J, Sullivan SA, Scarborough J, Wilke L, Pugh C, Bell RH, et al. Measuring competence in surgical training through assessment of surgical entrustable professional activities. *J Surg Educ*. 2018;75(6):1452–62. <https://doi.org/10.1016/j.jsurg.2018.05.004>
3. Williams RG, George BC, Bohnen JD, Dunnington GL, Fryer JP, Klamen DL, et al. A proposed blueprint for operative performance training, assessment, and certification. *Ann Surg*. 2021;273(4):701–8.
<https://doi.org/10.1097/SLA.0000000000004467>
4. Rela M, Price T. Review of the validity of DOPS as an assessment tool for the procedural skills of surgical trainees. *Ann R Coll Surg Engl*. 2023;105(7):599–606.
<https://doi.org/10.1308/rcsann.2022.0052>
5. Ghiasian L, Hadavandkhani A, Abdolalizadeh P, Janani L, Es'haghi A. Comparison of video-based observation and direct observation for assessing the operative performance of residents undergoing phacoemulsification training. *Indian J Ophthalmol*. 2021;69(3):574.
6. Hassanpour N, Chen R, Baikpour M, Moghimi S. Video observation of procedural skills for assessment of trabeculectomy performed by residents. *J Curr Ophthalmol*. 2016;28(2):61–4.
<https://doi.org/10.1016/j.joco.2016.03.003>
7. El-Sayed C, Yiu A, Burke J, Vaughan-Shaw P, Todd J, Lin P, et al. Measures of performance and proficiency in robotic assisted surgery: a systematic review. *J Robot Surg*. 2024;18(1):1–19.
<https://doi.org/10.1007/s11701-023-01756-y>
8. Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today*. 2013;43(3):271–5.
<https://doi.org/10.1007/s00595-012-0313-7>
9. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273–8.
<https://doi.org/10.1046/j.1365-2168.1997.02502.x>

10. Albadry M, Chen H, Settmacher U, Jiřík M, Dahmen U. A best-practice model for online teaching with feedback driven video analysis in surgical skills training. *J Surg Educ*. 2026;83(9):104062.
11. Wang X, Ji X. Sample size estimation in clinical research: from randomized controlled trials to observational studies. *Chest*. 2020;158(1 Suppl):S12–20.
<https://doi.org/10.1016/j.chest.2020.03.010>
12. Waqar SH. Students' perception of direct observation of procedural skills as workplace based assessment tool in general surgery. *Pak Armed Forces Med J*. 2016;66(5):731-7.
13. ElHawary H, Salimi A, Gorgy A, Fesdekjian L, Seal A, Gilardino MS. Medical student mentorship in surgery: lessons learnt and future directions. *J Surg Educ*. 2022;79(1):129-38.
14. Patil S, Allad A, Patil J, Patil P. Direct observation of procedural skills evaluation of suturing skills in surgical interns: a comprehensive analysis. *Asian J Med Sci*. 2024;15(7):102–7.
<https://doi.org/10.3126/ajms.v15i7.64979>
15. Balvardi S, Kammili A, Hanson M, Mueller C, Vassiliou M, Lee L, et al. The association between video-based assessment of intraoperative technical performance and patient outcomes: a systematic review. *Surg Endosc*. 2022;36(11):7938–48. <https://doi.org/10.1007/s00464-022-09296-6>
16. Nickel F, Hendrie JD, Stock C, Salama M, Preukschas AA, Senft JD, et al. Direct observation versus endoscopic video recording-based rating with the objective structured assessment of technical skills for training of laparoscopic cholecystectomy. *Eur Surg Res*. 2016;57(1-2):1–9.
<https://doi.org/10.1159/000444449>
17. Bhatti AM, Ahsin S, Mansoor S. Feasibility of video assessment of operative skills in FCPS vascular surgery. *Pak J Med Sci*. 2023;39(1):12–7. <https://doi.org/10.12669/pjms.39.1.6737>
18. McNally J, Wilson D, Jones EE, Shahisavandi N, King C, Coles C, et al. Video-based assessment of surgical skill in orthopaedic surgery: a technique guide to capturing critical surgical detail. *JB JS Open Access*. 2025;10(2):e24.00183.
19. Khaqan HA, Mahboob U, Haider MA, Buksh HM, Ur Rehman HA. Effectiveness of video observation of procedural skills in surgical performance of ophthalmology residents. *Pak J Med Health Sci*. 2022;16(3):169.
20. Gazis A, Schizas D, Kykalos S, Karaiskos P, Loukas C. Egocentric video analysis for automated assessment of open surgical skills via deep learning. *Int J Comput Assist Radiol Surg*. 2026;21(2):297-306.
21. Inamdar P, Hota PK, Undi M. Feasibility and effectiveness of direct observation of procedure skills (DOPS) in general surgery discipline: a pilot study. *Indian J Surg*. 2022;84(Suppl 1):109–114.
<https://doi.org/10.1007/s12262-021-02953-5>
22. Tantiplachiva K, Iramaneerat C, Lertbunnaphong T. Improving surgical skills with feedback: directly-observed versus video-recorded practice. *BMC Med Educ*. 2023;23(1):660.
<https://doi.org/10.1186/s12909-023-04635-0>
23. Yanik E, Ainam JP, Fu Y, Schwaitzberg S, Cavuoto L, De S. Video-based skill acquisition assessment in laparoscopic surgery using deep learning. *Glob Surg Educ J Assoc Surg Educ*. 2024;3(1):26.

<https://doi.org/10.1007/s44186-023-00223-4>

24. Liu D, Li Q, Jiang T, Wang Y, Miao R, Shan F, Li Z. Towards unified surgical skill assessment. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2021. p. 9522–9531. <https://doi.org/10.1109/CVPR46437.2021.00940>.
25. Benmansour M, Malti A, Jannin P. Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. *Int J Comput Assist Radiol Surg*. 2023;18(5):929–937. <https://doi.org/10.1007/s11548-022-02827-5>
26. Fatima B, Auqil Z, Zafar Z, Waheed MR, Bashir MK, Mujahid MUF. Unmasking the challenges: a qualitative study of barriers to training pediatric surgeons for minimally invasive surgery in Pakistan. *Pediatr Surg Int*. 2026;42(1):176. <https://doi.org/10.1007/s00383-026-06386-7>

Figures

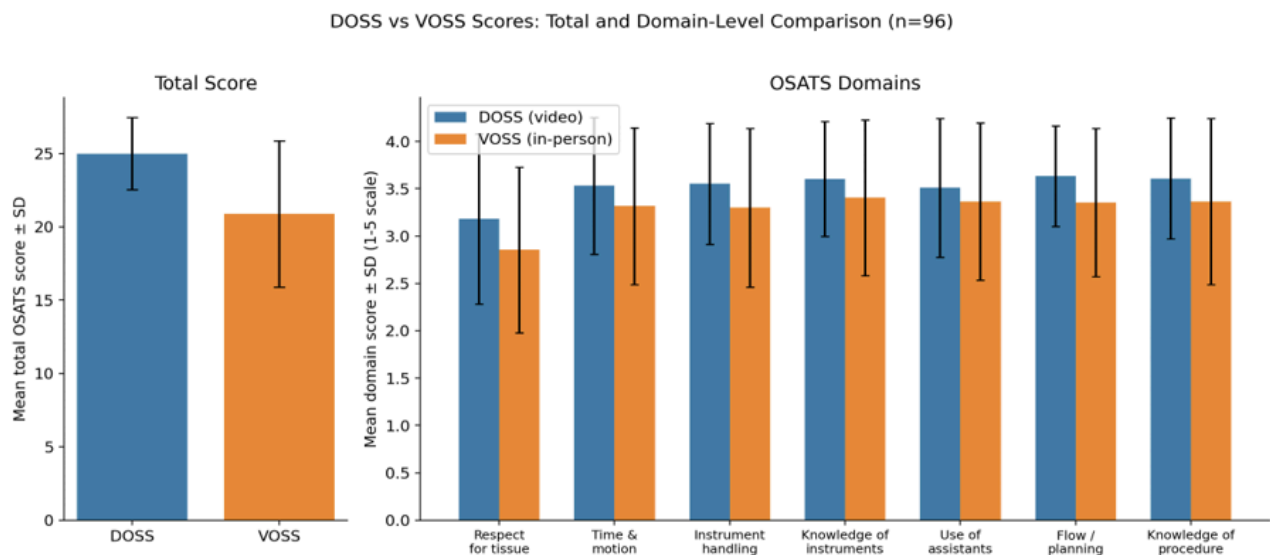


Figure 1

DOSS versus VOSS scores by domain and total

Figure 1. Mean DOSS and VOSS scores ($\pm$ SD) for the total OSATS score (left) and each of the seven OSATS domains (right), $n = 96$.

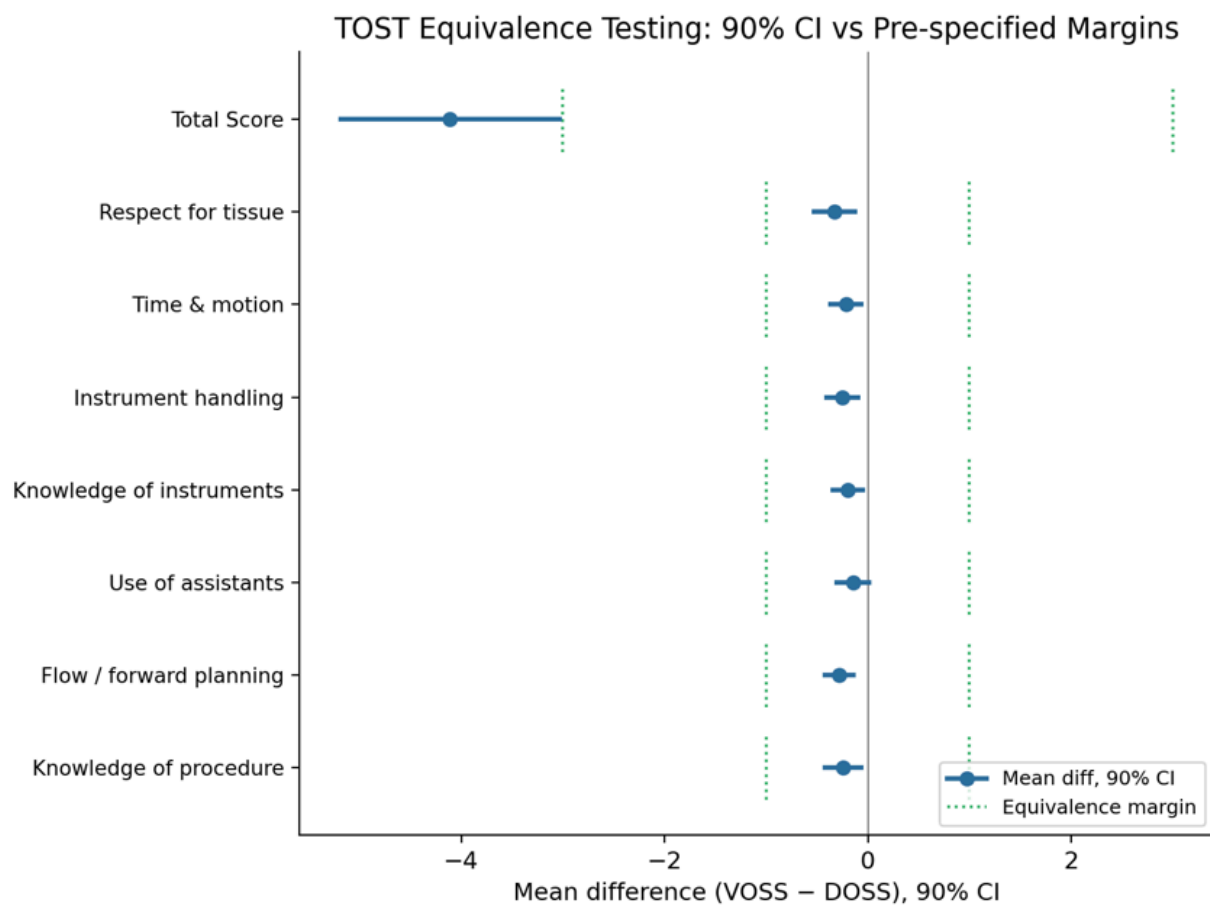


Figure 2

TOST equivalence plot

Figure 2. Mean difference (VOSS – DOSS) with 90% confidence interval for the total score and each OSATS domain, plotted against its prespecified equivalence margin (dotted green lines). Equivalence is supported when the entire CI falls within the margin.

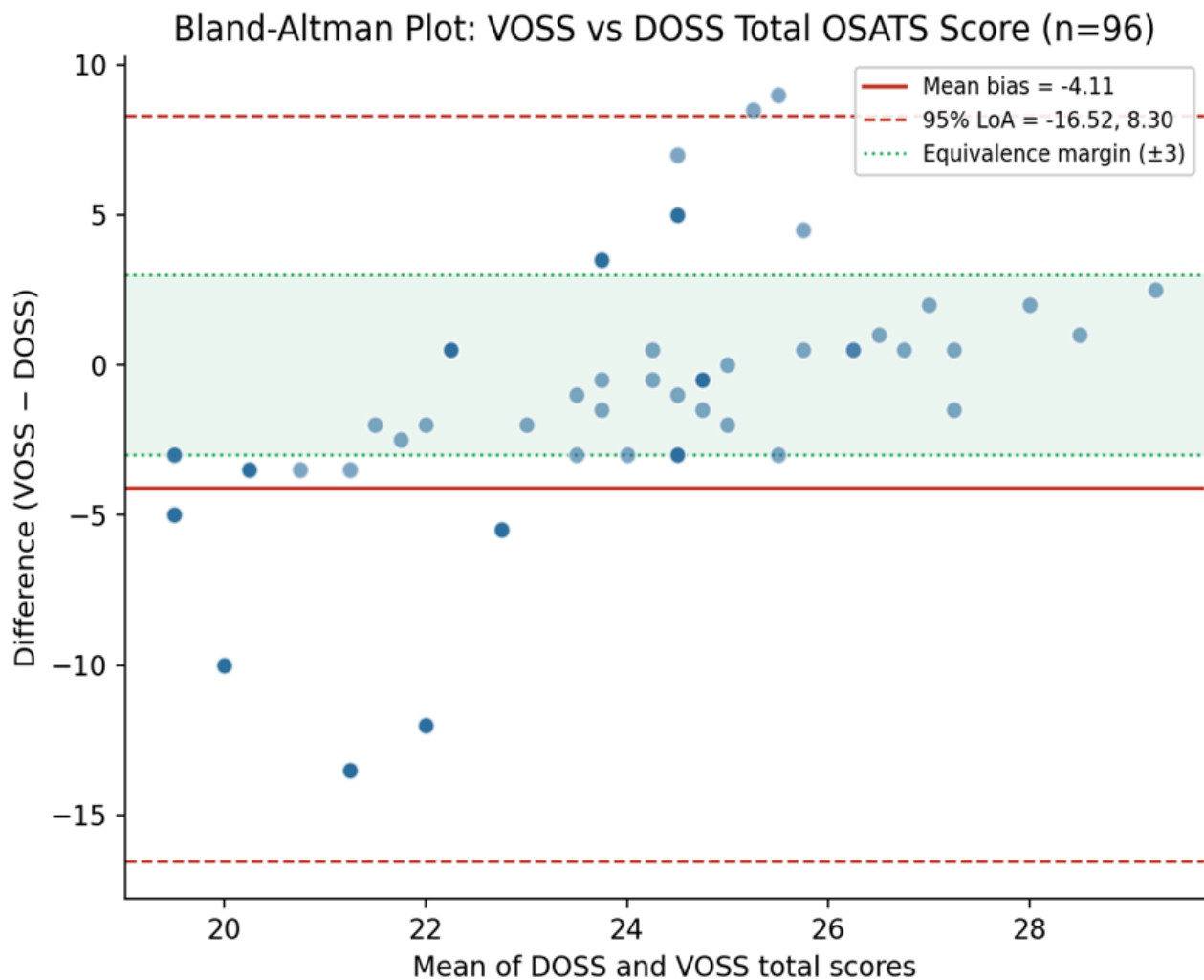


Figure 3

Bland-Altman plot, total OSATS score

Figure 3. Bland-Altman plot of the difference (VOSS - DOSS) against the pair mean for total OSATS score ($n = 96$). Solid red line, mean bias; dashed red lines, 95% limits of agreement; dotted green lines, prespecified equivalence margin of ± 3 points.