Inter-Rater Reliability of a Suture Assessment Tool

Khyatu Patel M.B.B.S.[1], Dillon Lundstrom M.D.[1], Elizabeth Husted M.D.[1], and Stephen K. Stacey D.O. FAAFP[1]

[1]Department of Family Medicine, Mayo Clinic Health System, La Crosse, WI 54601, USA.

Corresponding author: Dr. Stacey Stephen, D.O., Director of Osteopathic Education, Assistant Professor of Family Medicine, La Crosse-Mayo Family Medicine Residency, stacey.stephen@mayo.edu
Received: January 17, 2025, **Accepted:** June 6, 2025, **Published:** June 20, 2025
DOI: 10.5281/zenodo.15698667

Abstract

Introduction: Despite significant advancements in standardizing medical knowledge assessment, evaluating technical skills, such as suturing, has largely depended on subjective faculty evaluations, creating a gap in reliable, objective assessment. Addressing this need, Sundhagen et al. developed a suturing assessment tool, demonstrating promising reliability and validity in standardized settings with medical students. However, its performance remains untested when used by family medicine faculty to evaluate residents in a typical clinical environment. This pilot study aims to assess the inter-rater reliability of the modified Sundhagen tool in evaluating suturing skills among family medicine residents in real-world settings.
Methods: To assess inter-rater reliability, a modified version of the Sundhagen assessment tool, comprising eight yes-or-no questions on suturing skills, was provided to four family medicine faculty members in La Crosse, WI. Faculty members used this tool to evaluate five videos of family medicine residents performing simple sutures on chicken breast incisions, captured during routine educational activities without specific training on tool use. The videos, obtained from a prior residency study with participant consent, displayed only the residents' hands and suturing materials. Task completion scores were tallied, and inter-rater reliability was measured using ICC, with analysis by question.
Results: The ICC score was 0 (95% confidence interval -0.18 – 0.61), suggesting poor inter-rater reliability.
Conclusion: This pilot study did not demonstrate strong inter-rater reliability for the modified Sundhagen tool when applied to family medicine residents, highlighting the need for further refinement and assessment of additional validity components. As an essential preliminary step, these findings lay the groundwork for a larger, adequately powered trial that can more rigorously evaluate the tool's performance across a broader set of metrics, enhancing its applicability and reliability in family medicine residency training.

KEYWORDS: Suture assessment, surgical skills, reliability, family medicine, residency

1 Introduction

Despite advancements in standardizing medical knowledge assessment, technical skills evaluation in medical education often remains inconsistent, heavily reliant on subjective faculty assessments. These subjective assessments may provide inconsistent feedback to learners, creating the need for objective, reliable evaluation tools that can standardize procedural skill assessment [1].

Addressing this need, the Objective Structured Assessment of Technical Skills (OSATS) was developed to standardize surgical skill evaluation through a 7-dimensional rating system [2]. The rating system included respect for tissue, time and motion, instrument handling, knowledge of instruments, use of assistants, and specific procedures. Each domain is graded on a 5-point Likert scale. Since its development, the OSATS has been used to evaluate technical skills across multiple surgical specialities [2-6]. Despite successful adoption, concerns have risen regarding the ability of the OSATS to evaluate surgical outcomes [3]. There is also limited research on its use outside of surgical specialties.

To bridge this gap, Sundhagen et al. developed a new assessment tool for suturing skills specifically focused on medical students, including microsurgical and macrosurgical qualities [7]. Their assessment tool used eight yes-or-no questions addressing the following: proper use of instruments, knot tying, forceps use, suture use, needle insertion, suture placement, and time needed to complete the task. Their results showed promising reliability and validity with use in medical students.

Their original evaluation method involved grading by three board-certified plastic surgeons. While they reported excellent inter-rater reliability, whether this applies to training environments in other specialities remains undetermined. Building on our previous work—where an adaptive, real-time virtual work-

shop during the COVID-19 pandemic improved suture skills and confidence among family medicine residents—we now aim to assess the inter-rater reliability of the Sundhagen assessment tool in a family medicine context [8].

2 Methods

Subjects were recruited from the core faculty at the La Crosse-Mayo Family Medicine Residency Program in La Crosse, Wisconsin. They evaluated recordings of residents performing suture skills during routine experiences. These videos were previously obtained as part of a routine training workshop on laceration repair [8], and showed each resident performing a single simple interrupted suture tied with a square knot using an instrument tie. Equipment included needle holders, forceps, scissors, nylon sutures, and a chicken breast suture model. Only the hands are visible in all videos, and the videos contained no identifying information. Additionally, we did not include audio as that would have precluded blinding of trainees.

Six family medicine residency faculty were asked to assess the suturing videos utilizing the suture assessment tool using an email consent form. Four faculty members participated. They were each sent five videos showing a different subject performing the task. Raters were not given details about trainee characteristics other than the fact that they were family medicine residents.

In the original tool, the scoring formula was based on a previously published formula: cutoff time (seconds) – completion time (seconds) – (10 x sum of errors). As nine different variables were being evaluated, a maximum of nine errors could be committed. The completion time would be the total time to complete the task. A modified scoring system was used in this pilot study, with points awarded for an adequately performed skill without involving the time component. The main objective was to look for interrater reliability. The time taken from start to finish can be objectively determined, so the final score would differ from one reviewer to another based only on subjective components. The aim of the project was to evaluate the interrater reliability of subjectively determined by feedback component. The modification was removing the objective portion and keeping only the subjective portions as this study was meant specifically to evaluate reliability of the subjective portion. This was done to assess the reliability of individual subjective components of the score. A maximum of 9 points per evaluation was possible.

We used intra-class correlation (ICC) to assess the inter-rater reliability between our assessors. In addition, we used a two-way random effects model using absolute agreement as the relationship among raters and single as our unit of interest. Intraclass Correlation Coefficient (ICC) was used because the primary goal was to assess the level of agreement among multiple raters evaluating the same set of videos. By using the two-way model for absolute agreement with the single-rater unit, the analysis reflects how reliably an individual rater scores compared to others, which is directly relevant when determining the utility and consistency of the assessment tool across various evaluators.

Sundhagen assessment tool: Eight yes or no questions were included as stated below. In addition, amount of time needed to complete the two different tasks was measured in seconds (Table 1).

3 Results

Out of six faculty members invited to review each of five videos, four completed the reviews (66.7%). Our study showed poor agreement with an ICC of 0 (Table 2).

4 Discussion

The tool developed by Sundhagen et al. was chosen for this project due to its simplicity, which makes it well-suited for use in non-surgical specialties. However, while Sundhagen's initial study showed promising results in assessing the surgical skills of medical students, our study demonstrated limited inter-rater reliability when applied in a family medicine training context.

As a pilot study, the conclusions drawn from our results are limited. One factor contributing to the lower inter-rater reliability may have been the lack of standardization in video recordings. Our videos were obtained by learners using whatever recording devices they had available (typically cell phones). This approach was chosen to reflect a "real-world" application of the tool, although a more standardized recording setup may have improved scoring consistency. Other studies have used more standardized techniques, including multiple camera angles, which allowed for more precise assessment of specific suture characteristics, such as traction and needle positioning [9]. One of the raters also taught the laceration repair course, potentially influencing their interpretation of the videos which could result in rater's/observer bias.

Further investigation into the modified Sundhagen tool could examine additional forms of validity

Table 1: *Sundhagen assessment tool questions*

Q1. Did subject grab the needle with the instruments (and not with the fingers)
Q2. Did subject tie a correct squared knot
Q3. Did subject hold the forceps correctly
Q4. Did subject grab the suture with the instruments in a correct fashion (in a way that does not potentially lead to suture breakage)
Q5. Did subject penetrate the foam suture pad with a 90 degrees angle
Q6. Did subject manage the suture without tangling the ends in the knot
Q7. Did subject damage the foam suture pad
Q8. Did subject make a parallel suture (equal length from the wound edge and equal depth on both sides

Table 2: *Family Medicine Faculty Ratings of Resident Suture Performance Using the Modified Sundhagen Scoring System*

Videos	Grader 1	Grader 2	Grader 3	Grader 4
Video 1	8	8	8	8
Video 2	8	6	8	7
Video 3	8	8	4	7
Video 4	8	8	7	7
Video 5	8	8	4	7

testing. Future studies might focus on evaluating content validity, construct validity, concurrent validity, inter-item reliability, and test-retest reliability to build a more comprehensive evidence base for the tool's applicability. Additionally, the development and use of standardized "control" videos—both "perfect" and "imperfect" examples—that can help calibrate individual raters and assess grading difficulty is also worth exploring. This approach could offer valuable insight into rater variability and improve the reliability of the assessment tool. Investigating whether rater training is needed before tool use may also enhance consistency. Analyzing individual question responses and scores could identify which items contribute to lower ICC values, guiding refinement of the tool. Furthermore, examining how different questions perform in family medicine versus surgical settings may support the development of a more reliable, context-specific assessment tool.

5 Conclusion

Our study did not demonstrate sufficient inter-rater reliability of the modified Sundhagen tool within a family medicine residency setting. However, further investigation into this tool's applicability could help establish it as an effective method for assessing this important procedural skill.

Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Data Availability

Not applicable.

Conflicts of Interest

The authors declare that they have no competing interests.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Author Contributions

6 References

1. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. Acad Med. 1996;71(12):1363-1365. doi:10.1097/00001888-199612000-00023
2. Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. Am J Obstet Gynecol. 2006;195(2):617-623. doi:10.1016/j.ajog.2006.05.032
3. Anderson DD, Long S, Thomas GW, Putnam MD, Bechtold JE, Karam MD. Objective Structured Assessments of Technical Skills (OSATS) Does Not Assess the Quality of the Surgical Result Effectively. Clin Orthop Relat Res. 2016;474(4):874-881. doi:10.1007/s11999-015-4603-4
4. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84(2):273-278. doi:10.1046/j.1365-2168.1997.02502.x
5. Kramp KH, van Det MJ, Hoff C, Lamme B, Veeger NJ, Pierie JP. Validity and reliability of global operative assessment of laparoscopic skills (GOALS) in novice trainees performing a laparoscopic cholecystectomy. J Surg Educ. 2015;72(2):351-358. doi:10.1016/j.jsurg.2014.08.006
6. Asif H, McInnis C, Dang F, et al. Objective Structured Assessment of technical skill (OSATS) in the Surgical Skills and Technology Elective Program (SSTEP): Comparison of peer and expert raters. Am J Surg. 2022;223(2):276-279. doi:10.1016/j.amjsurg.2021.03.064
7. Sundhagen HP, Almeland SK, Hansson E. Development and validation of a new assessment tool for suturing skills in medical students. Eur J Plast Surg. 2018;41(2):207-216. doi:10.1007/s00238-017-1378-8
8. Stacey SK, Boswell CL, Cowan KK, et al. Remote Synchronous Laceration Repair Instruction With Summary Feedback. PRiMER. 2023;7:19. Published 2023 Jun 29. doi:10.22454/PRiMER.2023.863182
9. Brisson BA, Dobberstein R, Monteith G, Jones-Bitton A. Excellent Agreement of In-Person Scoring versus Scoring of Digitally Recorded Simulated Suture Skills Examination. J Vet Med Educ. Published online July 7, 2022. doi:10.3138/jvme-2021-0164