PRiMER
peer-reviewed reports in medical education research
research brief

Interrater Reliability of a Suture Assessment Tool in Family Medicine Training

Khyati Patel, MMBS | Carmen Dargel, MD | Ricky A. Aguayo-Ortega, DO | Christopher L. Boswell, MD | Stephen K. Stacey, DO

PRiMER. 2026;10:2.

Published: 2/2/2026 | DOI: 10.22454/PRiMER.2026.615323

Abstract

Introduction: Procedural competence is essential to clinical practice, yet assessment of technical skills such as suturing remains highly variable and often relies on subjective evaluations. While tools with limited validity evidence exist in surgical specialties, their use in family medicine is limited. This study evaluated the interrater reliability of a suture assessment tool developed by Sundhagen et al when used by family medicine faculty to evaluate the simple interrupted suturing skills of family medicine residents.

Methods: In this cross-sectional reliability study, 15 core faculty members from four family medicine residency programs were recruited to evaluate the simple interrupted suturing skills of 20 family medicine residents using a modified suture assessment tool developed by Sundhagen et al. Intraclass correlation coefficients were used to evaluate interrater reliability. Light's κ was calculated for each question.

Results: A total of 15 evaluators assessed 20 videos of resident suturing performance. The Light's κ estimates ranged from 0.11 to 0.50, indicating slight to moderate agreement. None of the items reached the threshold for substantial agreement (κ≥0.61).

Conclusions: The modified Sundhagen suture assessment tool demonstrated fair to moderate reliability when used to evaluate family medicine residents. While the level of agreement limits the tool's effectiveness for standardized summative assessment, the tool may still offer value as a structured framework for formative feedback. Future studies may improve reliability through refined rater training, behaviorally anchored scoring criteria, and the use of visual exemplars.

Introduction

Despite the critical role that procedural competence plays in patient outcomes, assessment of technical skills such as suturing remains highly variable.[1,2] While assessments for medical knowledge have become largely objective and reproducible, the evaluation of technical skills continues to depend primarily on subjective faculty impressions.[3,4] In an era where patient safety and clinical outcomes are increasingly tied to the precision of hands-on care, the lack of standardized, validated tools to assess suturing ability represents a serious gap in medical education.[1,2,4]

Family physicians frequently perform minor procedures, including suturing, as part of routine practice.[5] Yet,

procedural skill assessment in family medicine remains highly variable, often relying on informal faculty observation rather than structured, objective tools.[6,7] While educators have developed tools aimed at standardizing surgical skill evaluation across surgical specialties, these tools have not been evaluated in family medicine contexts.[2-12] For example, the objective structured assessment of technical skills (OSATS) is a performance-based examination aimed at assessing the clinical competence of surgical residents.[2,8,10] However, its use remains limited by concerns over predictive validity and sparse application outside traditional surgical settings.[3,4] To address these gaps, Sundhagen et al developed an assessment tool specifically for evaluating suturing skills in medical students. Their study demonstrated promising reliability and validity in medical student populations.[2]

Developing reliable approaches for evaluating suturing performance could enhance both feedback quality and resident skill development. Therefore, the purpose of this study was to evaluate the interrater reliability of the Sundhagen suture assessment tool when applied to family medicine residents. By doing so, we aimed to determine whether the tool may serve as a feasible, objective framework for procedural skills assessment in family medicine residency training.

Methods

This study aimed to evaluate interrater reliability (IRR) of the suture assessment tool developed by Sundhagen et al when used by family medicine residency faculty to assess family medicine residents. The study was deemed exempt by the Mayo Clinic Institutional Review Board.

Core faculty from four family medicine residency programs at a single multispecialty institution in the Midwest region of the United States evaluated videos of family medicine residents performing simple interrupted sutures. Participants were eligible for inclusion if they were active core faculty at one of the participating family medicine residency programs and provided consent.

Faculty evaluators independently reviewed 20 de-identified videos of family medicine residents performing simple interrupted sutures. Videos were obtained from routine procedure training sessions at a single residency program, with family medicine residents filmed individually under standardized lighting, consistent camera positioning, and high-resolution recording conditions (Figure 1). To preserve anonymity, only gloved hands were visible, and family medicine residents provided written informed consent for video capture.

Evaluators used a modified version of the Sundhagen suture assessment tool to rate each performance. The original tool includes both subjective assessments and an objective time-based component; however, for this study, only the subjective criteria were assessed, and the time factor was excluded. Evaluators were anonymized using assigned identification numbers. They each received a standard instructional script that outlined the study purpose, explained that the modified Sundhagen tool should be applied exactly as written, and provided logistical guidance for completing ratings in REDCap. This minimal-instruction approach was intentional to evaluate the response process in real-world conditions.[13]

IRR was assessed using Light's κ for each yes/no checklist item, calculated with the kappam.light function in R (R Foundation). This chance-corrected measure computes the mean Cohen's κ across all possible pairs of raters. A sample of 15 raters evaluating 20 videos provided greater than 80% power (two-sided α=.05) to detect a difference of 0.2 in reliability estimates. κ values are interpreted as follows: 0 reflects no agreement; 0.01 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, near-perfect agreement.[14]

Results

A total of 15 faculty reviewers consented to participate, each rating 20 videos. The Light's κ estimates ranged from 0.11 to 0.50, indicating fair/moderate IRR for most questions (Table 1). None of the items reached the threshold for substantial agreement (κ≥0.61).

Discussion

This study aimed to evaluate the IRR of a modified version of the Sundhagen suture assessment tool for use in assessing simple interrupted suturing by family medicine residents. The tool demonstrated fair to moderate IRR, indicating substantial variability in scoring across evaluators. While this variability may limit the tool's effectiveness as a standardized summative assessment, the tool may still offer value as a structured framework for formative feedback and resident development—pending further refinement before broader implementation.

This study offered several methodological strengths, including the use of prerecorded standardized videos, prior rater training, and an adequately powered design to estimate IRR. However, the study was limited to a single institution, which may affect generalizability. Refining the training protocol with more detailed calibration procedures may have improved the IRR, though such enhancements could also limit generalizability by deviating from typical real-world implementation.

Several factors may have contributed to the limited IRR observed, including ambiguous language in the original assessment tool, which was not modified for this study. The tool may benefit from clearer scoring anchors to guide evaluators in distinguishing between adjacent rating categories. Inconsistent scoring may also have resulted from vague criteria for items such as "correct forceps use" or "damage to the suture model." Enhancing the tool by providing visual exemplars and explicit definitions may also improve consistency. Future studies also may explore improving IRR by providing structured calibration sessions using annotated videos and consensus discussions.[15]

Conclusions

This study found that a structured assessment tool for evaluating simple interrupted suturing by family medicine residents demonstrated fair to moderate IRR, reflecting variability in scoring across evaluators. These findings suggest that while the tool may be useful for guiding formative feedback for family medicine residents, the tool shows limited utility for summative or standardized assessments. Further refinement—including behaviorally anchored scoring criteria, standardized rater training, and integration into structured formats—may improve reliability and support the tool's broader application in family medicine residency education.

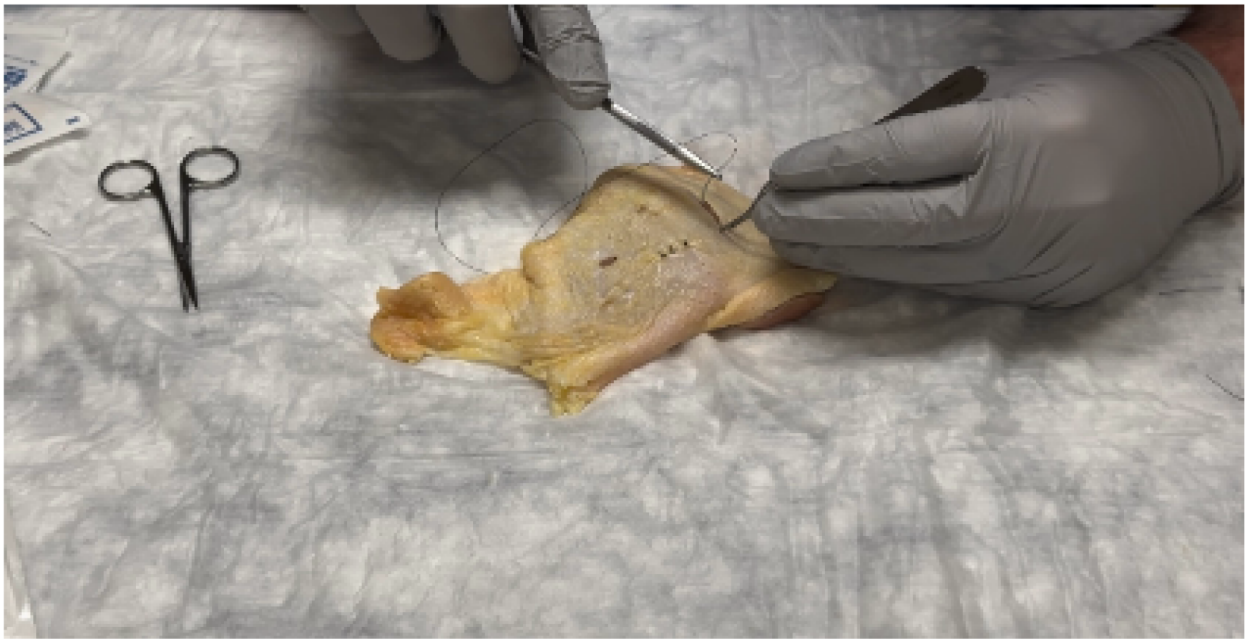
Tables and Figures

Table 1. Interrater Reliability Statistics

Question	Light's κ	Interpretation of agreement
1. Did the subject grab the needle with the instruments (and not with the fingers)?	0.26	Fair
2. Did the subject tie a correct squared knot?	0.37	Fair
3. Did the subject hold the forceps correctly?	0.24	Fair
4. Did the subject grab the suture with the instruments in a correct fraction (in a way that does not potentially lead to suture breakage)?	0.17	None to slight
5. Did the subject penetrate the suture model with 90 degree angle?	0.34	Fair
6. Did the subject manage the suture without tangling the ends in the knot?	0.50	Moderate
7. Did the subject damage the suture model?	0.11	None to slight
8. Did the subject make a parallel suture (equal length from the wound edge and equal depth on both sides)?	N/A	N/A

Abbreviation: N/A, not applicable

Figure 1. Example Frame From Video Recording



Acknowledgments

The authors sincerely thank all the Mayo family medicine core faculty and residents who participated. We also extend our thanks to Natile Averkamp, MS, for her contributions as a biostatistician. Without their involvement, this study would not have been possible.

Presentations: Mayo Clinic Southwest Wisconsin Annual Research Day, May 20, 2025, Mayo Clinic Health System, La Crosse, WI.

Conflicts of Interest: The authors have no conflicts of interest to declare.

Corresponding Author

Stephen K. Stacey, DO

Department of Family Medicine, Mayo Clinic Health System, La Crosse, WI

Stacey.stephen@mayo.edu

Author Affiliations

Khyati Patel, MMBS - Family Medicine Residency, Department of Family Medicine, Mayo Clinic Health System, La Crosse, WI

Carmen Dargel, MD - Department of Family Medicine, Mayo Clinic Health System, La Crosse, WI

Ricky A. Aguayo-Ortega, DO - Family Medicine Residency, Department of Family Medicine, Mayo Clinic Health System, La Crosse, WI

Christopher L. Boswell, MD - Rochester and Kasson Residency Program, Department of Family Medicine, Mayo Clinic Health System, Kasson, MN

Stephen K. Stacey, DO - Department of Family Medicine, Mayo Clinic Health System, La Crosse, WI

References

1. Patel K, Lundstrom D, Husted E, Stacey S. Inter-rater reliability of a suture assessment tool. *Intermountain Journal of Translational Medicine*. 2025;2(1). doi:10.5281/zenodo.15698667

2. 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

3. 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

4. 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

5. Newman AR, Heidelbaugh JJ, Klemenhagen K, Michelfelder AJ, Power DV, Hougas JE. Current procedural practices of family medicine teaching physicians. *Fam Med*. 2024;56(3):156-162. doi:10.22454/FamMed.2024.197714

6. Garcia-Rodriguez JA, Dickinson JA, Perez G, et al. Procedural knowledge and skills of residents entering Canadian family medicine programs in Alberta. *Fam Med*. 2018;50(1):10-21. doi: 10.22454/FamMed.2018.968199

7. Kedian T, Gussak L, Savageau JA, Cohrssen A, Abramson I, Everard K, Dobbie A. An ounce of prevention: how are we managing the early assessment of residents' clinical skills? A CERA study. *Fam Med*. 2012 Nov-Dec;44(10):723-6.

8. 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):1,363-1,365. doi:10.1097/00001888-199612000-00023

9. 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

10. MacRae H, Regehr G, Leadbetter W, Reznick RK. A comprehensive examination for senior surgical residents.

Am J Surg. 2000;179(3):190-193. doi:10.1016/S0002-9610(00)00304-4

11. 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

12. 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):617-621. doi:10.1016/j.ajog.2006.05.032

13. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575. doi:10.1111/medu.12678

14. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282. doi:10.11613/BM.2012.031

15. Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract*. 2007;12(2):239-260. doi:10.1007/s10459-006-9043-1