Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability and Feasibility of a Multi-Source, Competency-Based Evaluation System for Surgical Sub-Internship Students: A Single-Institution Pilot Study

Anthony Onde Morada

anthony.omorada@gmail.com

Geisinger Wyoming Valley Medical Center https://orcid.org/0000-0002-0428-6558

Jessica L Becker

Geisinger Wyoming Valley Medical Center

Michael J Furey

Geisinger Wyoming Valley Medical Center

Christian Hailey Summa

: Geisinger Wyoming Valley Medical Center

Kristen R Richards

Geisinger Wyoming Valley Medical Center

Ajmal Baray

Geisinger Wyoming Valley Medical Center

Megan L Gordon

Geisinger Wyoming Valley Medical Center

Akshay Patel

Geisinger Wyoming Valley Medical Center

Rebecca Michelle Jordan

Geisinger Wyoming Valley Medical Center

Joseph P Bannon

Geisinger Community Medical Center

Alexandra Falvo

Geisinger Community Medical Center

Research Article

Keywords: sub-internship, surgical education, multi-source feedback, inter-rater reliability, competency-based assessment, residency selection, workplace-based assessment

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10029588/v1

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

Abstract

Background

Sub-internship (Sub-I) rotations are high-stakes components of the surgical residency selection process, yet most programs rely on informal, single-source assessments. We sought to develop and pilot a structured, multi-stakeholder, competency-based evaluation instrument and assess its inter-rater reliability, feasibility, and discriminant validity.

Methods

A six-domain evaluation instrument was developed using behaviorally-anchored 5-point Likert scales aligned with Accreditation Council for Graduate Medical Education (ACGME) core competencies. Attending surgeons, residents, and clinical staff submitted evaluations via a mobile-accessible digital form over a 5-month pilot period at a single academic general surgery residency program. Inter-rater reliability was assessed using the two-way random-effects intraclass correlation coefficient [ICC(2,1)] per the Shrout-Fleiss framework, interpreted per Koo and Li guidelines. Discriminant validity was evaluated through analysis of flagged performance concerns.

Results

Eight sub-interns received 24 evaluations (mean 3.0 per student) from attending physicians (17%), residents (79%), and clinical staff (4%) across inpatient, operative, and outpatient settings. Median completion time was 3.2 minutes. ICC values demonstrated excellent reliability for overall intern-readiness (0.93; 95% CI: 0.79–0.995) and technical skills (0.92; 95% CI: 0.65–1.00); good reliability for initiative and preparation (0.79; 95% CI: 0.48–0.98) and communication (0.75; 95% CI: 0.43–0.98); and moderate reliability for clinical knowledge (0.66; 95% CI: 0.30–0.97). Performance concerns were flagged in 17% of evaluations. No significant score differences were identified between attending and resident evaluators (all $p > 0.05$).

Conclusions

A brief, multi-source, competency-based Sub-I evaluation instrument achieved strong inter-rater reliability for the domains most central to residency selection, with minimal evaluator burden. Larger multi-institutional studies are needed to define optimal evaluator thresholds and correlate Sub-I assessment scores with downstream residency performance and program fit.

Introduction

Sub-internship (Sub-I) rotations represent a defining interaction in surgical residency selection. Often described as high-stakes “audition” periods, Sub-I experiences directly shape program directors’ assessments of candidates and, by extension, rank list positioning [1]. Despite the ultimate weight of their influence in the resident selection process, the evaluative practices supporting Sub-I assessment in general surgery remain largely unstructured, non-standardized, and inconsistently applied across programs [2, 3]. Studies conducted in an attempt to provide a Sub-I evaluation process with structure have found that patient care and professionalism were far more important to raters than raw technical skill and knowledge [4]. The average length of a Sub-I rotation runs about 4 weeks and this short time period adds another variable for program directors and students to battle with, as students will ultimately spend a short time period with each supervising attending. Many students aim to secure a Sub-I letter of recommendation, while program directors assess their character and personality to determine program fit [2, 5].

The consequences of this gap are meaningful. Without validated evaluation frameworks, assessors default to global impressions susceptible to availability bias, recency effects, and differential faculty exposure across clinical settings. Students receive limited formative feedback to guide professional development and program directors lack objective data to substantiate selection decisions or identify at-risk candidates early in the rotation. For students from underrepresented backgrounds or those with limited prior relationship with program faculty, the absence of structured assessment may amplify existing inequities inherent to the selection process. Institutions commonly have large scale DEI initiative, however individualized career focused support tends to be missing for those underrepresented in medicine (URiM) [6].

Competency-based medical education (CBME), codified through the Accreditation Council for Graduate Medical Education (ACGME) core competency framework, has fundamentally restructured graduate medical education over the past two decades [7]. The resulting portfolio of validated workplace-based assessment (WBA) tools—including the Objective Structured Assessment of Technical Skills (OSATS), Mini-Clinical Evaluation Exercise (Mini-CEX), and 360-degree multi-source feedback instruments—demonstrates reliable, context-sensitive evaluation across diverse clinical environments [8–11]. These tools share key design principles: behaviorally-anchored structured scales, multi-rater perspectives, and prospective validity evidence.

The application of these principles to Sub-I assessment has received limited systematic attention. Multi-source feedback, in particular, offers compelling advantages for the Sub-I context. Especially in the context of surgical education, where time to complete surveys is limited and evaluator participation is thus often unreliable, there exists a need for an evaluation tool that offers high inter-rater reliability and requires fewer overall evaluations to yield meaningful feedback. Additionally, whereas standard evaluation tools generally seek only attending and resident physician feedback, expanding the breadth of evaluator subtypes to include other clinical staff represents an untapped assessment resource [11].

The purpose of this study was to: (1) develop a structured, multi-stakeholder, competency-based Sub-I evaluation instrument grounded in validated WBA design principles; (2) pilot the instrument prospectively over a 5-month implementation period at a single academic general surgery residency program; and (3) assess the instrument's inter-rater reliability, implementation feasibility, and capacity to discriminate between levels of student performance.

Methods

Study Design and Setting

This was a prospective quality improvement (QI) pilot study evaluating a novel Sub-I assessment instrument implemented at a single academic general surgery residency program. The study was reviewed by the institutional review board and designated as Quality Improvement / Exempt from full review. Data were collected over a continuous 5-month implementation window, which corresponded with the Sub-I rotation season.

Instrument Development

With the assistance of Claude Haiku 4.5 [12], the evaluation tool was developed by the institution's Sub-Internship Interview and Experience Committee following a structured review of the surgical and medical education WBA literature. Five competency domains, mapped to ACGME core competencies, were evaluated across behaviorally-anchored 5-point Likert items (1 = Significant concerns; 2 = Below expectations; 3 = Meets expectations; 4 = Above expectations; 5 = Exceptional) [13]. The five domains evaluated on our survey were: 1) application of clinical knowledge to patient care, 2) competence with technical or procedural skills, 3) communication effectiveness with the team, 4) preparation and initiative-taking, 5) professional behavior under stress or when receiving feedback.

The first survey domain assessed, the application of clinical knowledge to patient care (1), maps to ACGME medical knowledge and patient care competencies [13]. The assessment of this domain is supported by prior use of the mini-CEX, which demonstrates high internal consistency ($\alpha = 0.92–0.97$) for assessing clinical reasoning and diagnostic accuracy through brief structured encounters [10]. The second survey domain, competence with technical or procedural skills also maps to the ACGME patient care and procedure skills core competency [13]. The assessment of this domain is supported by previous studies evaluating OSATS, using global rating scales for technical skills assessment, with an inter-rater reliability of 0.78–0.85 [8, 9]. The third survey domain, communication effectiveness with the team maps to the ACGME competency interpersonal and communication skills [13]. The assessment of this domain is supported by the use of the Communication Assessment Tool (CAT) which demonstrates strong inter-rater reliability (intraclass correlation coefficient, ICC, 0.82–0.83) across observer types, and incorporation of 390-degree feedback for assessment of teamwork (ICC 0.57–0.83) [11, 14]. Preparation and initiative-taking (4), connects to the ACGME competency of practice-based learning and

improvement [13]. Its assessment is supported by professionalism indicators including accountability and service orientation documented in multi-source feedback validation studies [11]. Professional behavior under stress or when receiving feedback (5), correlates directly with the ACGME professionalism competency, and its assessment is supported by previous studies evaluating the reliability and validity of using multi-source feedback in nonclinical and clinical domains [11, 13].

A 5-point Likert scale was utilized as global rating scales demonstrate superior psychometric properties and outperform checklists for reliability and validity with behavioral anchoring [8]. A 6th survey question using the previously stated 5-point Likert scale was also included to obtain a global assessment to gauge the student's overall performance or readiness for intern year. The use of global ratings for assessment methods has been found to outperform task-specific checklists for both reliability and validity [8]. An N/A option was included for each Likert item to prevent forced ratings when evaluators lacked direct observational experience in a specific domain, improving data quality without penalizing partial evaluators [10]. Evaluators also documented the primary clinical setting of their interaction (operating room, inpatient floor, clinic, or office/non-patient care) to enable context-specific performance analysis.

Three open-ended narrative items supplemented the quantitative ratings: (1) the student's greatest observed strength, (2) one specific area for improvement, and (3) a three-level concern flag — No concerns, Minor concerns, or Major concerns — with a conditional free-text field requiring elaboration when Minor or Major concerns were selected. The tiered concern structure was designed to triage safety-relevant observations to the residency selection committee without requiring evaluators to render explicit pass/fail judgments.

The instrument was deployed via Microsoft Forms and optimized for mobile access to minimize potential barriers to completion in clinical environments.

Evaluator Recruitment and Data Collection

Attending surgeons, residents (of all post-graduate year levels), and clinical staff (registered nurses, advanced practice providers, and operating room technicians) with direct supervisory or team-based contact with Sub-I students were solicited as evaluators. We chose to design a multi-source evaluation tool as professional colleagues require fewer evaluators than trainees for equivalent reliability, as demonstrated in a previous study reaching a high reliability coefficient of ($\alpha > 0.8$) with 5 professional staff evaluations vs 23 medical student evaluations [15]. Evaluator type was classified against a pre-defined roster of faculty and trainees using name matching. All students rotating on the surgical service during the pilot period were included.

Raw data were exported from Microsoft Forms in CSV format and processed in R (version 4.x; R Foundation for Statistical Computing, Vienna, Austria) using the *dplyr*, *tidyr*, *psych*, and *irr* packages. Test entries and blank responses were removed prior to analysis. Students and evaluators were assigned

random anonymous identifier codes prior to any analysis intended for research dissemination. Completion time was calculated as the interval between form start and submission timestamps.

Statistical Analysis

Inter-rater reliability was assessed using the two-way random-effects single-measure intraclass correlation coefficient [ICC(2,1)], consistent with the framework proposed by Shrout and Fleiss [16]. This model was selected because both students and raters represent random samples drawn from larger populations and students were not assessed by a fixed panel of evaluators, precluding a fixed-effects model. The single-measure form was used to estimate the reliability of individual evaluations, providing a conservative estimate of instrument reproducibility. ICC values and 95% confidence intervals were calculated using the *psych* package. Analyses were restricted to students with three or more evaluations per domain to ensure minimum reliability of ICC estimates.

ICC values were interpreted as: poor (< 0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (> 0.90) per published guidelines [17]. Thresholds of 0.60 and 0.75 were used in graphical display to delineate the moderate-to-good and good-to-excellent transitions.

Score distributions were characterized using medians and interquartile ranges (IQR) given the ordinal nature of Likert-scale data. Evaluator type comparisons (attending vs. resident) were made using independent-samples t-tests; a p-value < 0.05 was considered statistically significant. Evaluator threshold analysis was attempted using an ICC-based matrix approach; sparse evaluator overlap across the small cohort precluded empirical threshold determination, and literature-based benchmarks were used in its place [18].

Results

Participant and Evaluator Characteristics

Eight sub-interns were evaluated over the 5-month implementation period, generating 24 total evaluations (mean 3.0 evaluations per student; range: 1–8). Evaluators comprised attending physicians (n = 4; 17%), residents (n = 19; 79%), and clinical staff (n = 1; 4%). Clinical settings were not mutually exclusive; evaluations were most frequently completed in the inpatient floor setting (n = 18; 75%), followed by the operating room (n = 17; 71%), clinic (n = 9; 38%), and office/non-patient care (n = 2; 8%). Median completion time was 3.2 minutes (IQR: 1.9–5.3 min). Participant characteristics are summarized in Table 1.

Table 1
Participant and Evaluator Characteristics

Students	**N = 24**	**%**
Students evaluated	8	—
Evaluations per student, mean (range)	3.0 (1–8)	—
Evaluator type		
Attending physicians	4	17%
Residents	19	79%
Clinical staff (RN / APP / OR tech)	1	4%
Clinical setting *(not mutually exclusive)*		
Operating room	17	71%
Inpatient floor	18	75%
Clinic	9	38%
Office / non-patient care	2	8%
Implementation		
Study duration	5 months	—
Median completion time, min [IQR]	3.2 [1.9, 5.3]	—
Clinical settings are not mutually exclusive; evaluators could indicate multiple interaction contexts.		

Inter-Rater Reliability

ICC(2,1) values and 95% confidence intervals were calculable for all six competency domains (Table 2; Fig. 1). Overall intern-readiness demonstrated excellent reliability (ICC = 0.93; 95% CI: 0.79–0.995), as did technical skills (ICC = 0.92; 95% CI: 0.65–1.00). Good reliability was demonstrated for initiative and preparation (ICC = 0.79; 95% CI: 0.48–0.98) and communication (ICC = 0.75; 95% CI: 0.43–0.98). Clinical knowledge demonstrated moderate reliability (ICC = 0.66; 95% CI: 0.30–0.97). Professionalism demonstrated the lowest ICC of the six domains (ICC = 0.47; 95% CI: 0.12–0.94), categorized as poor per standard thresholds, though this finding reflects ceiling effects rather than instrument failure.

Table 2. Inter-Rater Reliability by Competency Domain — ICC(2,1) Results

Domain	ICC	95% CI	Interpretation
Overall intern-readiness	0.93	0.79–0.995	Excellent
Technical skills	0.92	0.65–1.00	Excellent
Initiative & preparation	0.79	0.48–0.98	Good
Communication	0.75	0.43–0.98	Good
Clinical knowledge	0.66	0.30–0.97	Moderate
Professionalism	0.47	0.12–0.94	Poor*

The professionalism domain ICC should be interpreted in the context of ceiling effects (see Discussion). ICC: intraclass correlation coefficient; CI: confidence interval. Interpretation thresholds per Koo and Li [17]: poor < 0.50, moderate 0.50–0.75, good 0.75–0.90, excellent > 0.90.

Competency Score Distributions

Students demonstrated strong performance across competency domains (Table 3). Professionalism received the highest ratings (86% scoring ≥ 4), followed closely by initiative and preparation. Overall intern-readiness assessments yielded median scores of 5 (IQR 4–5), with 78% of evaluations rating students as highly prepared (≥ 4 out of 5). Technical skills demonstrated the widest score distribution and lowest median among all six domains, consistent with expected developmental variability in procedural competency acquisition during sub-internship training. Clinical knowledge showed ceiling clustering, with the narrowest score range across all domains, suggesting evaluators perceived consistent baseline cognitive competency in this group.

Table 3. Competency Score Distributions by Domain and Evaluator Type

Domain	n	Median [IQR]	% ≥ 4	Attending Mean (SD)	Resident Mean (SD)	p-value
Overall intern-readiness	23	5 [4–5]	78%	2.75 (1.71)	4.44 (0.98)	0.14
Clinical knowledge	21	4 [4–5]	76%	4.00 (1.00)	4.06 (0.73)	0.94
Technical skills	20	4 [3.75–4.25]	75%	3.00 (1.83)	3.94 (0.93)	0.39
Communication	24	4.5 [4–5]	75%	3.25 (1.26)	4.32 (1.00)	0.19
Initiative & preparation	23	5 [4–5]	83%	3.00 (1.41)	4.61 (0.78)	0.10
Professionalism	22	5 [4.25–5]	86%	4.33 (1.15)	4.56 (0.98)	0.78

p-value for independent-samples t-test comparing attending vs. resident mean scores. IQR: interquartile range; SD: standard deviation.

Performance Concern Identification

Of 24 evaluations, 83% (n = 20) were marked “No concerns,” 12% (n = 3) “Minor concerns,” and 4% (n = 1) “Major concerns.” At the student level, 1 of 8 students (12%) received at least one evaluation with a concern flag, accounting for approximately 17% of all evaluations. The conditional narrative field was completed in all flagged evaluations, providing specific behavioral descriptions available for selection committee review.

Evaluator Type Comparisons

No statistically significant differences in mean domain scores were identified between attending physicians and residents across any of the six competency domains (all $p > 0.05$). Mean scores and standard deviations by evaluator type are presented in Table 3.

Discussion

This pilot study demonstrates that a brief, multi-source, competency-based evaluation instrument can achieve robust inter-rater reliability in the surgical Sub-I setting. The tool required a median of 3.2 minutes to complete; was feasible across inpatient, operative, and outpatient environments; and achieved ICC values in the good-to-excellent range for the two domains most central to residency selection decision-making — overall intern-readiness and technical skill.

The excellent ICC values for overall intern-readiness (0.93) and technical skills (0.92) are particularly notable in context. These domains most directly address the judgment that informs program director decision-making: whether a candidate will function safely and effectively as a first-year surgical resident. For comparison, the OSATS — the most extensively validated WBA tool in surgical education — reports

inter-rater reliability of 0.78–0.85 using global rating scales [8, 9]. The Mini-CEX demonstrates internal consistency of α = 0.92–0.97 for clinical reasoning domains [10]. Our findings are consistent with, and in several domains exceed, these established benchmarks. This is particularly meaningful given the small sample size of the pilot, which would be expected to widen confidence intervals and, if anything, underestimate the instrument's long-run reliability properties.

The relatively lower ICC for professionalism (0.47) warrants interpretation in context rather than dismissal as an instrument flaw. Ceiling effects are a well-documented challenge in behavioral assessment when rater variance is constrained by uniformly high performance [19, 20]. In our pilot, professionalism received the highest median ratings and the lowest score dispersion of any domain. When the majority of students are rated similarly, the statistical variance necessary for ICC computation is inherently restricted — producing lower estimates that reflect score homogeneity, not genuine inter-rater disagreement. Multi-source feedback studies in graduate medical education have similarly reported professionalism domain ICCs ranging from 0.47 to 0.83 depending on whether ceiling clustering is present [11]. The formative value of this domain — captured through the narrative items and tiered concern flag — substantially exceeds what the ICC statistic alone can represent.

The multi-stakeholder design produced evaluations weighted heavily toward resident contributors (79%), reflecting the operational reality of surgical training environments where trainees maintain considerably more patient-contact hours with Sub-I students compared to attending surgeons. This distribution is consistent with 360-degree assessment models in graduate medical education and represents a practical strength of the design: resident evaluators contribute ecologically valid, longitudinal performance observations that a single end-of-rotation attending assessment cannot replicate [11]. Critically, no statistically significant score differences were identified between attending and resident evaluators across any domain, supporting the validity of multi-source score aggregation and suggesting that evaluator type did not introduce systematic bias. This finding aligns with evidence from multi-source feedback research demonstrating that professional healthcare colleagues require substantially fewer raters than trainee-based systems to achieve equivalent reliability thresholds — an estimated 5 vs. 23 evaluators for ICC ≥ 0.80 — making this design practically sustainable in busy surgical programs [11].

The system demonstrated clear discriminant validity: 17% of evaluations were flagged with performance concerns, and the tiered structure (no/minor/major) enabled proportionate, actionable triage without requiring evaluators to render explicit binary pass/fail judgments. This sensitivity to subthreshold performance is a critical instrument property for any tool embedded in residency selection, where failing to identify a struggling candidate carries downstream consequences for both the student and the program. The conditional narrative field in all flagged evaluations provides specific behavioral evidence for selection committee deliberation — a qualitative richness that complements the quantitative ICC data.

Sample size constraints prevented empirical determination of an optimal evaluator threshold in this pilot. The data matrix — 24 evaluations distributed across 8 students and multiple evaluators with minimal

rater overlap — did not permit matrix-based ICC threshold simulation as described by Bonett [18]. Based on converging evidence from multi-source feedback research, a target of 4–5 evaluations per student is supported for achieving acceptable reliability thresholds among professional rater groups [11]. Our pilot achieved a mean of 3.0 evaluations per student, indicating that modest scale-up of evaluator participation in future cycles would likely reach this benchmark.

This study has several limitations that should be acknowledged. The sample is small — 8 students, 24 evaluations — and reflects a single institution over a 5-month pilot window, limiting generalizability. The heavy reliance on resident evaluators (79%) may introduce evaluator-type clustering not fully captured in the current analysis framework. Confidence intervals for several domain ICC estimates are wide, reflecting small-sample uncertainty rather than instrument instability. Empirical evaluator threshold determination was not feasible. Longitudinal correlation with residency match outcomes and post-match intern performance was not assessed, representing the most consequential gap in the validation chain that this pilot is designed to initiate. Another limitation to acknowledge is that subinterns were not provided a set of objectives or list of expectations prior to starting their rotations. The formal documentation and provision of such objectives would not only enable students to hone their performance, but would also create an opportunity for raters to make direct comparisons among Sub-I students, knowing that all students had been given the same pre-rotation objectives.

In addition to development and distribution of Sub-I rotation objectives, future directions include expansion to a multi-institutional cohort to enable definitive evaluator threshold determination through statistical simulation; prospective correlation of Sub-I evaluation scores with intern-year performance ratings; integration with Electronic Residency Application Service (ERAS) applicant data for predictive validity analysis; and examination of inter-setting reliability across operative, inpatient, and outpatient environments. The instrument's design as a continuous quality improvement infrastructure — not solely a research instrument — means that longitudinal data accumulation across implementation cycles will progressively strengthen the evidence base with each cohort of Sub-I students assessed.

Conclusions

A brief, multi-source, competency-based evaluation instrument for surgical sub-interns demonstrated excellent inter-rater reliability for overall intern-readiness and technical skills, good reliability for communication and initiative and preparation, moderate reliability for clinical knowledge, and adequate performance across all six domains — with a median evaluator burden of 3.2 minutes per evaluation. The instrument successfully identified students with performance concerns, confirming discriminant validity. These pilot findings support the feasibility and psychometric soundness of structured, evidence-based Sub-I evaluation in general surgery. Scaling this framework has the potential to enhance transparency, equity, and rigor in the residency selection process while simultaneously providing formative, multi-perspective feedback that supports student development throughout the Sub-I rotation.

Declarations

Data Availability Statement

The data in this study was collected via Microsoft Forms and included sub-intern name and photograph to assist attending surgeons, residents, and clinical staff with proper identification when completing the form. The six competency domains and Likert items used are discussed in the Instrument Development section. De-identified data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

1. Aziz H, Khan S, Rocque B, et al. Selecting the Next Generation of Surgeons: General Surgery Program Directors and Coordinators Perspective on USMLE Changes and Holistic Approach. World J Surg. 2021;45:3258–65. https://doi.org/10.1007/s00268-021-06261-7.
2. Lindeman BM, Lipsett PA, Alseidi A, Lidor AO. Medical student subinternships in surgery: characterization and needs assessment. Am J Surg. 2013;205:175–81. https://doi.org/10.1016/j.amjsurg.2012.10.008.
3. Sidlow R. The structure and content of the medical subinternship: A national survey. J Gen Intern Med. 2001;16:550–3. https://doi.org/10.1046/j.1525-1497.2001.016008550.x.
4. Rhodin KE, Thornton SW, Leraas HJ, et al. Roles, Responsibilities, and Expectations of Medical Students on Surgical Subinternships: A Modified Delphi Consensus Study. Acad Med. 2024;99:874–83. https://doi.org/10.1097/ACM.0000000000005685.
5. Naples R, Shin TH, French JC, Lipman JM. Beyond Medical Knowledge and Patient Care: A Program Director's Perspective for the Role of General Surgery Subinternships. J Surg Educ. 2021;78:83–90. https://doi.org/10.1016/j.jsurg.2020.06.028.
6. Rasic G, Jung S, Dechert A T, et al. A national survey of diversity, equity, and inclusion efforts in general surgery residency programs. Global Surg Educ. 2024;3:102. https://doi.org/10.1007/s44186-024-00298-7.
7. Swing SR. The ACGME outcome project: retrospective and prospective. Med Teach. 2007;29:648–54. https://doi.org/10.1080/01421590701392903.
8. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents: OBJECTIVE STRUCTURED ASSESSMENT OF TECHNICAL SKILL. Br J Surg. 1997;84:273–8. https://doi.org/10.1046/j.1365-2168.1997.02502.x.
9. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. Acad Med. 1996;71:1363–5. https://doi.org/10.1097/00001888-199612000-00023.
10. Chen Y-Y, Chiu Y-C, Chu T-S, et al. Is the rating result reliable? A new approach to respond to a medical trainee's concerns about the reliability of Mini-CEX assessment. J Formos Med Assoc. 2022;121:943–9. https://doi.org/10.1016/j.jfma.2021.07.005.

11. Donnon T, Al Ansari A, Al Alawi S, Violato C. (2014) The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review. Academic Medicine 89:511–516.
https://doi.org/10.1097
/ACM.0000000000000147 (2026)Release notes | Claude Help Center.
https://support.claude.com/en/articles/12138966-release-notes. Accessed 12 June 2026.
12. Association of American Medical Colleges AA of C of OM. Foundational competencies for undergraduate medical education. AAMC, AACOM, and ACGME Washington, DC; 2024.
13. Yuen JK, Kelley AS, Gelfman LP, et al. Development and Validation of the ACP-CAT for Assessing the Quality of Advance Care Planning Communication. J Pain Symptom Manag. 2020;59:1–e83. https://doi.org/10.1016/j.jpainsymman.2019.09.001.
14. Massagli TL, Carline JD. Reliability of a 360-Degree Evaluation to Assess Resident Competence. Am J Phys Med Rehabil. 2007;86:845–52. https://doi.org/10.1097/PHM.0b013e318151ff5a.
15. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86:420–8. https://doi.org/10.1037/0033-2909.86.2.420.
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15:155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
17. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002;21:1331–5. https://doi.org/10.1002/sim.1108.
18. Williams RG, Klamen DL, White CB, et al. Tracking Development of Clinical Reasoning Ability Across Five Medical Schools Using a Progress Test. Acad Med. 2011;86:1148–54. https://doi.org/10.1097/ACM.0b013e318226631b3.
19. Albanese MA. Commentary: Measurement and Interpretation Challenges in Comparing Student Performance Outcomes From Different Medical Schools. Acad Med. 2011;86:1073–5. https://doi.org/10.1097/ACM.0b013e318226340c.

Figures

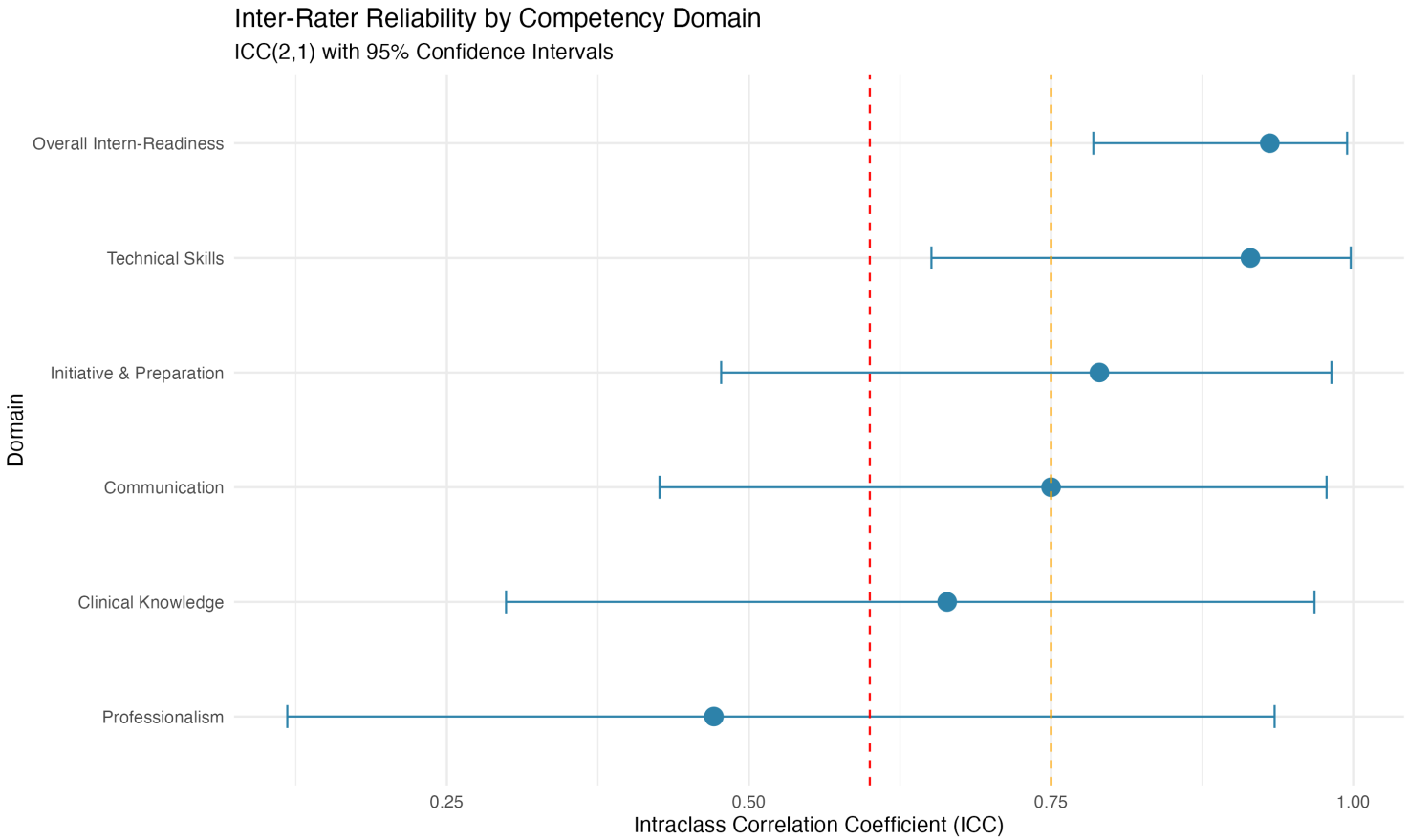


Figure 1

Inter-rater reliability by competency domain. Forest plot displaying ICC(2,1) point estimates with 95% confidence intervals for each of the six competency domains. Dashed vertical lines indicate thresholds for moderate (0.60) and good (0.75) reliability per Koo and Li [17].