Published in final edited form as:
Adv Health Sci Educ Theory Pract. 2015 March ; 20(1): 181–191. doi:10.1007/s10459-014-9519-3.

Verification of accurate technical insight: a prerequisite for self-directed surgical training

Yinin Hu, M.D., **Helen Kim**, **Adela Mahmutovic**, **Joanna Choi**, **Ivy Le**, and **Sara Rasmussen, M.D., Ph.D.**
University of Virginia School of Medicine, Department of Surgery

Abstract

BACKGROUND—Simulation-based surgical skills training during preclinical education is a persistent challenge due to time constraints of trainees and instructors alike. Self-directed practice is resource-efficient and flexible; however, insight into technical proficiency among trainees is often lacking. The purpose of this study is to prospectively assess the accuracy of self-assessments among medical students learning basic surgical suturing.

METHODS—Over seven weekly practice sessions, preclinical medical students performed serial repetitions of a simulation-based suturing task under one-on-one observation by one of four trainers. Following each task repetition, self- and trainer-assessments were performed using a 36-point weighted checklist of technical standards developed *a priori* by expert consensus. Upon study completion, agreement between self- and trainer-assessments was measured using weighted Cohen's kappa coefficients.

RESULTS—Twenty-nine medical students each performed a median of 25 suture task repetitions (IQR 21.5–28). Self-assessments tended to overestimate proficiency during the first tertile of practice attempts. Agreement between self- and trainer-assessments improved with experience, such that the weighted kappa statistics for the two-handed and instrument ties were greater than 0.81 after 18 to 21 task attempts.

CONCLUSIONS—Inexperienced trainees frequently overestimate technical proficiency through self-assessments. However, this bias diminishes with repetitive practice. Only after trainees have attained the capacity to accurately self-assess can effective self-directed learning take place.

Keywords

Medical Student Education; Self-directed learning; Self-assessment; Simulation training; Surgical Education; Suturing

INTRODUCTION

Effective self-directed learning is perhaps the most valuable competency acquired during medical school, and is one that is indispensable for sustained certification (S. H. Miller.

Corresponding Author: Sara Rasmussen, M.D., Ph.D., PO Box 800709, Charlottesville, VA 22908-0679, skr3f@virginia.edu, P: 1-(434) 982-2796, F: 1-(434) 243-0036.

Conflicts of Interest: None

2005; A. Bandura. 1977). Accurate self-assessment is a prerequisite to self-directed learning for technical skills (GO Grow. 1991; O. Safir et al. 2013). This factor is particularly relevant to simulation-based training due to the substantial opportunity costs incurred by retaining clinicians as evaluators. Thus, as the role of preclinical simulation expands due to the increasing scrutiny of quality and safety metrics, there is a greater need than ever to investigate trainee self-assessment in technical skills.

Simulation training among medical students is most effective with repetitive practice and proper self-reflection on performance (J. M. Sargeant et al. 2009; N. Taffinder et al. 1998). However, self-assessments may not accurately depict proficiency levels, and are hindered by absence of meaningful benchmarks and insufficient understanding of technical nuances (P. A. Lipsett et al. 2011; T. R. Eubanks et al. 1999). This lack of metacognition, or the ability to judge one's own performance, is most evident among lower-performing novices, and may improve with experience and skill level (J. Krueger and R. A. Mueller. 2002; J. Kruger and D. Dunning. 1999). However, there is a paucity of research aimed at identifying thresholds of experience beyond which self-assessments become reliable (T. R. Anthoney. 1986; M. J. Gordon. 1992). This inattention is misguided, as progress in this territory lends greater validity to self-directed learning.

We aimed to examine longitudinally the development of technical insight using a model that addresses a basic surgical skill: knot-tying. The purpose of this study was to analyze trends in agreement between self- and trainer-assessments in a simulation-based suturing workshop. We hypothesized that accuracy of self-assessments would increase with repetition, that there exists an experience threshold beyond which self- and trainer-assessments have comparable validity, and that this threshold varies across different technical skills.

METHODS

First- and second-year medical students without prior clinical experience were voluntarily enrolled within a longitudinal suturing skills module. This module belonged to a larger study spanning weekly practice sessions over three months that also included modules for intubation and central venous catheterization. For an expected total participation time of between 10–15 hours, participants each received a stipend of $100. The suturing skills module began with an orientation session with instruction on instrument handling and basic knot-tying technique for two-handed, one-handed, and instrument-tie knots. Subsequently, during independently scheduled weekly practice sessions, participants repetitively performed a suture task which involved three consecutive figure-of-eight sutures—one secured with a two-handed knot, one with a one-handed knot, and one with an instrument tie. The three sutures were performed with 2-0 silk suture, a needle driver, and forceps, and approximated a 1.0 cm-wide simulated wound.

Scoring criteria for the task were derived from task-specific checklist items within preexisting suturing Objective Structured Assessment of Technical Skills (OSATS) (P. D. van Hove et al. 2010; J. A. Martin et al. 1997; J. G. Chipman and C. C. Schmitz. 2009). To better capture technical nuances of the suturing task used in the study, additional error

definitions and penalty weights were defined *a priori* by consensus among surgical faculty. For analytic purposes, the composite score is the combined weighted checklist outcome (maximum 36 points), while knot-specific sub-scores are comprised of items relevant to each knot-tying technique (Table 1); for example, the two-handed knot sub-score has a maximum of 9 points. Participants were not formally trained to use the weighted checklist, however, all checklist components were described as part of the orientation session. Participants were informed of a 5-minute time constraint guideline for proficiency; however, task attempts violating the time criterion were not excluded from self- and trainer-assessment using the weighted checklist.

Trainers were comprised of four undergraduate students, each in their third year of premedical education. Trainers each underwent five hours of individual instruction by surgical faculty and staff prior to study initiation. Following instruction, all four trainers demonstrated technical mastery of the suturing task as assessed by a surgical faculty member. To ensure consistent scoring criteria, the four trainers then concurrently observed ten consecutive task attempts by a single participant and submitted assessments using the objective technical scoring checklist. Using this method, Cohen's weighted Kappa for inter-rater agreement was 0.866.

Each participant received one-on-one oversight from one of the four experienced trainers during every practice session. trainers were assigned on a weekly rotating schedule such that all participants were exposed to each trainer for an equal number of sessions. During practice sessions, self- and trainer-assessments were performed simultaneously using the same scoring criteria immediately following every task attempt. Trainer assessments were based on direct observation of technique and manual tension-check for air knots following task completion.

Trainer-assessment results and technical feedback were not provided to participants until after the corresponding self-assessment was completed. Each practice session incorporated a minimum of three suturing tasks with corresponding assessments. Self-motivated practice outside of practice sessions was neither encouraged nor discouraged, but sutures and instruments were not provided to participants outside of practice sessions. Following the completion of seven sessions, all participants underwent post-test evaluation by a member of the surgical faculty using the same weighted checklist. During post-testing, the suturing task was performed three times, and the top two scores were averaged and recorded. Assuming that technical skill does not vary significantly between the last practice session and the post-test, this faculty evaluation was considered a final quality check on technical proficiency as well as self- and instructor-assessment accuracy.

Each participant's practice volume was divided into tertiles, and the average difference between self- and trainer-assessment scores (SA-TA) within each tertile was determined for each participant. Checklist item-specific SA-TA differences were normalized as percent error ([SA-TA]/[max points]) and averaged across the study population; this provides a relative measure of the likelihood of self-assessment overestimation for each item. To determine if self-assessment accuracy is related to technical proficiency, the first tertile's results were compared between high- and low-performing subgroups using the Student's *t*-

test due to normal data distribution. To determine changes in the directionality of assessment differences, the average SA-TA across all participants was calculated for every practice attempt, and its relationship with practice experience was assessed using univariate linear regression. Cohen's weighted kappa coefficients were calculated for every practice attempt across all participants to assess changes in the level of agreement between self- and trainer-assessments as practice volume increased. Lastly, we compared how closely the SA and the TA on participants' last practice attempt approximated post-test faculty assessments, also using Cohen's weighted kappa. We considered weighted kappa greater than or equal to 0.81, 0.61, and 0.41 to be indicative of excellent, good, and moderate agreement, respectively (J. R. Landis and G. G. Koch. 1977; P. Brennan and A. Silman. 1992). All data were analyzed using SAS statistical software (version 9.3; SAS Institute, Inc). This study was approved by the University of Virginia Institutional Review Board (IRB-SBS protocol #2013-0246-00).

RESULTS

Thirty medical student participants were enrolled, of which 29 completed all requisite practice sessions and were included for analysis. Participants were comprised of 10 first-year students and 19 second-year students and 69% (20/29) were male. Median age was 21 years (IQR 20–22). Over seven sessions, participants on average each completed 25 suturing task attempts (IQR 22.25 – 28) and spent 2.15 practice hours dedicated to suturing (IQR 2.04 – 2.56). Upon study completion, median proficiency score by faculty post-test evaluation was 35 out of 36 (IQR 33–36).

The difference between self- and trainer-assessments (SA-TA) was calculated for every task attempt performed by each participant. The overall range of SA-TA for the study population was from −10 to +16, with the majority of values between −2 and +4. The distribution of average SA-TA for each participant is shown in Figure 1. Participants tended to overestimate performance during the first tertile of practice attempts (Figure 1A), with improvements in accuracy as experience increased (Figure 1C). During the first tertile of attempts, the checklist items most susceptible to self-assessment over-estimation were: suture-drop during one-handed ties (+12.6% average error), air knots on one-handed (+10.1%), instrument (+9.9%) and two-handed ties (+7.4%), and appropriate suture tail length on instrument tie (+6.9%). Assessing the trend in average SA-TA across all participants, it is notable that average SA-TA—representing overestimation—significantly decreased with experience ($p = 0.031$). Variance in SA-TA increased beyond 30 attempts, as relatively few participants were able to complete more than 30 practice attempts in the allotted sessions (Figure 2).

Participants were subdivided based on initial proficiency (by average TA score over the first 3 attempts) and post-test proficiency (by faculty post-test results), relative to median. Under subgroup analysis of first tertile practice data, participants who were below-median in initial proficiency overestimated composite scores and one-handed tie sub-scores to a greater extent than high-proficiency participants (Table 2). There were no statistically-significant associations between post-test performance and early SA accuracy.

Agreement between SA and TA across all participants was assessed for each task attempt using Cohen's weighted kappa. For the composite task score, SA accuracy increased with experience such that excellent agreement (weighted kappa ≥ 0.81) between SA and TA was achieved after 22 attempts (Figure 3). Knot-tying technique sub-scores were also analyzed individually. For the two-handed and instrument ties, agreement between SA and TA was excellent after 18 to 21 attempts. Interestingly, although the gap between SA and TA for the one-handed tie also narrowed with experience, the trend was substantially more gradual, and only moderate agreement was achieved after 30 attempts. The level of agreement between participants' last-attempt SA's and the faculty post-test assessments was 0.74 for the composite score, comparing favorably to the level of agreement between last-attempt TA's and faculty assessments (0.68).

DISCUSSION

The present study is the first to utilize serial self- and trainer-assessments to monitor the acquisition of suturing and knot-tying proficiency. Our data reinforce the concept that novices frequently lack self-awareness and overestimate technical skill, particularly early on in training. However, by analyzing the gaps between self- and trainer-assessments, we describe a threshold of experience beyond which a trainee may be considered sufficiently perceptive for accurate self-assessment. Furthermore, we demonstrate variability in the acquisition of technical insight that is related to the training tasks themselves. As teaching paradigms continue to explore roles for self-directed learning, aptitude with using assessment metrics must be verified before novices can be relied upon to self-assess.

Simulation-based surgical assessments most commonly focus on laparoscopic skills (R. Aggarwal et al. 2006; J. R. Korndorffer Jr et al. 2005; A. M. Pearson et al. 2002). By setting objective criteria for error and speed, laparoscopic simulation modules have attained excellent internal and external validity, culminating in the Fundamentals of Laparoscopic Surgery (FLS) curriculum (S. A. Fraser et al. 2003; M. C. Vassiliou et al. 2006; A. M. Derossis et al. 1998). However, junior surgical residents' initial operative experiences are comprised almost entirely of open surgeries. Because surgical exposure during medical school varies greatly (D. K. Nakayama and A. Steiber. 1990), "boot camps" that train graduating students and junior residents in open surgical techniques are increasingly common (M. E. Klingensmith and L. M. Brunt. 2010; L. M. Brunt et al. 2008; R. A. Stewart et al. 2007). Self-assessments in these boot camps generally involve either reviews of video recordings or repetitive OSATS-type evaluations (Y. Hu et al. 2013; J. MacDonald et al. 2003; R. Brydges et al. 2010), and aim to improve metacognition, morale, motivation, and communication (C. Hildebrand et al. 2009; J. Stewart et al. 2000; M. J. Gordon. 1992). Furthermore, self-assessments may reduce opportunity cost, as retaining surgical faculty as evaluators takes away from clinical revenue. Offsetting these benefits, however, is the fact that self-assessments are frequently inaccurate, particularly among low-performing and inexperienced trainees (K. W. Gow. 2013; P. A. Lipsett et al. 2011; T. R. Anthoney. 1986).

There are a number of methods by which self-assessments may attain improved accuracy. First is the adoption of objective scoring criteria with clearly-defined endpoints. Comprised of task-specific checklists and a global rating score, OSATS have high internal and external

validity (P. D. van Hove et al. 2010; J. A. Martin et al. 1997). Although global rating scores are more pervasive due to correlation with experience level and ease of implementation (G. J. Xeroulis et al. 2007; L. S. Mandel et al. 2005), they are less granular and more reliant upon clinical expertise than their checklist counterparts. Conversely, a major drawback of extensive checklists is the considerable time resource necessary for repetitive scoring (D. J. Scott et al. 2007). Augmenting self-assessments with motion analysis metrics may be one way to reduce personnel costs (G. J. Xeroulis et al. 2007; V. Datta et al. 2006; R. Brydges et al. 2009). Innovative metrics based on motion analysis and wound closure measurements may correlate with expert assessments (R. Brydges et al. 2009; V. Datta et al. 2006). Unfortunately, these metrics tend to focus on only a few aspects of proper surgical technique, and may be associated with prohibitive fixed costs. To balance these factors, we implemented a pragmatic weighted checklist that is founded on clinical expertise and has proven inter-rater reliability (0.866). Because this checklist is short (between 4–5 items for each knot), even relatively small differences in scoring reflect clinically-relevant variations in technique.

Perhaps the most intuitive way by which self-assessments gain validity is through trainee experience. To date, the literature addressing longitudinal changes in self-assessment with increasing experience remains divided. One early study comparing sequential self- and faculty-evaluations among medical students found that agreement actually decreased over a six-year curriculum (L. Arnold et al. 1985). Similar trends were highlighted by Fitzgerald and colleagues, who noted a decrease in correlation between self- and expert-assessments as students became exposed to clinical environments during the third year of training (J. T. Fitzgerald et al. 2003). Within surgery, MacDonald and colleagues showed that correlation between trainee error estimation and computer simulator data improved with repetition for FLS-type tasks (J. MacDonald et al. 2003). On the other hand, Ward and colleagues concluded that exposure to benchmarking videos as a means of augmenting trainee experience did not further improve the accuracy of self-assessments among residents performing laparoscopic fundoplication (M. Ward et al. 2003).

By requiring assessments after every task attempt, this study demonstrates the convergence between self- and experienced trainer-assessments with increasing experience. We provide evidence that self-assessments for two-handed and instrument tie techniques are accurate and valid after a threshold of 18 to 22 repetitions. Establishing this threshold for a diverse range of technical skills is worthwhile, as it would lend greater credibility to self-directed learning curricula. The data also indicate that convergence between self- and trainer-assessments is more gradual for the one-handed technique. One possible explanation is that technical insight may develop more slowly for tasks that are more advanced. Because of the intricacies involved in the one-handed tie, the two-handed technique is almost universally taught first. Both the two-handed and instrument tie techniques involve deliberate hand-crossing movements. By mentally tracking these deliberate movements, trainees may be more conscientious of knot quality. Applying this concept to clinical situations, it is apparent that self-directed training for tasks that require subtle technical skills will require careful scrutiny, and should be audited with periodic expert evaluations.

This study has several limitations. First, despite pragmatic advantages, employing undergraduate students as trainers may reduce the clinical relevancy of one-on-one instruction. All trainers underwent formal evaluation by a surgical faculty member and were judged to be skilled in the simulated task, however, bedside validation was infeasible. Similarly, although excellent inter-rater reliability was captured *a priori*, this was assessed through simultaneous observation of sequential attempts by a single participant in simulation rather than varied attempts across multiple participants in a clinical setting. By showing that last-attempt self- and trainer-assessments enjoy good agreement with faculty-administered post-test results, we bolster the legitimacy of these assessments. However, because self- and trainer-assessments were not performed during the faculty post-test, this remains an imperfect comparative standard. Second, speed was not factored into assessments. Recognizing that participants are entry-level trainees, the study protocol was designed to advocate form over pace. Furthermore, as time measurements do not vary between self- and trainer-assessments, the inclusion of speed as a criterion would have biased outcomes by artificially increasing agreement. Finally, our results cannot reveal the causative factor which contributed most to self-assessment accuracy. Over the study curriculum, participants gained practice experience, technical skill, and exposure to the assessment tool. In reality, all three components are integral to self-directed learning in surgery, and we advocate for further, focused research aimed at dissecting the relative contributions of each component.

In summary, repetitive practice has positive effects not only on technical proficiency but trainee perceptiveness as well. Importantly, technical insight develops at different rates for different techniques. Once trainee insight is calibrated over an initial period of concurrent trainer assessments, self-directed training protocols for surgical skills can be both resource-efficient and valid.

Acknowledgments

Funding source: University of Virginia Academy of Distinguished Educators: Undergraduate Medical Education Research and Innovation Grant

References

1. Miller SH. American Board of Medical Specialties and repositioning for excellence in lifelong learning: maintenance of certification. J Contin Educ Health Prof. 2005; 25(3):151–156. [PubMed: 16173049]
2. Bandura A. Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev. 1977; 84(2): 191–215. [PubMed: 847061]
3. Grow G. Teaching Learners to be Self-Directed. 1991; 41(3):125–149.
4. Safir O, Williams CK, Dubrowski A, Backstein D, Carnahan H. Self-directed practice schedule enhances learning of suturing skills. Can J Surg. 2013; 56(6):E142–7. [PubMed: 24284153]
5. Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. Adv Health Sci Educ Theory Pract. 2009; 14(3):399–410. [PubMed: 18528777]
6. Taffinder N, Sutton C, Fishwick RJ, McManus IC, Darzi A. Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. Stud Health Technol Inform. 1998; 50:124–130. [PubMed: 10180527]

7. Lipsett PA, Harris I, Downing S. Resident self-other assessor agreement: influence of assessor, competency, and performance level. Arch Surg. 2011; 146(8):901–906. [PubMed: 21844433]
8. Eubanks TR, Clements RH, Pohl D, Williams N, Schaad DC, Horgan S, Pellegrini C. An objective scoring system for laparoscopic cholecystectomy. J Am Coll Surg. 1999; 189(6):566–574. [PubMed: 10589593]
9. Krueger J, Mueller RA. Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. J Pers Soc Psychol. 2002; 82(2):180–188. [PubMed: 11831408]
10. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol. 1999; 77(6):1121–1134. [PubMed: 10626367]
11. Anthoney TR. A discrepancy in objective and subjective measures of knowledge: do some medical students with learning problems delude themselves? Med Educ. 1986; 20(1):17–22. [PubMed: 3951375]
12. Gordon MJ. Self-assessment programs and their implications for health professions training. Acad Med. 1992; 67(10):672–679. [PubMed: 1388532]
13. van Hove PD, Tuijthof GJ, Verdaasdonk EG, Stassen LP, Dankelman J. Objective assessment of technical surgical skills. Br J Surg. 2010; 97(7):972–987. [PubMed: 20632260]
14. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997; 84(2): 273–278. [PubMed: 9052454]
15. Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. J Am Coll Surg. 2009; 209(3): 364–370. e2. [PubMed: 19717041]
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–174. [PubMed: 843571]
17. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ. 1992; 304(6840):1491–1494. [PubMed: 1611375]
18. Aggarwal R, Hance J, Undre S, Ratnasothy J, Moorthy K, Chang A, Darzi A. Training junior operative residents in laparoscopic suturing skills is feasible and efficacious. Surgery. 2006; 139(6):729–734. [PubMed: 16782426]
19. Korndorffer JR Jr, Dunne JB, Sierra R, Stefanidis D, Touchard CL, Scott DJ. Simulator training for laparoscopic suturing using performance goals translates to the operating room. J Am Coll Surg. 2005; 201(1):23–29. [PubMed: 15978440]
20. Pearson AM, Gallagher AG, Rosser JC, Satava RM. Evaluation of structured and quantitative training methods for teaching intracorporeal knot tying. Surg Endosc. 2002; 16(1):130–137. [PubMed: 11961623]
21. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc. 2003; 17(6): 964–967. [PubMed: 12658417]
22. Vassiliou MC, Ghitulescu GA, Feldman LS, Stanbridge D, Leffondre K, Sigman HH, Fried GM. The MISTELS program to measure technical skill in laparoscopic surgery : evidence for reliability. Surg Endosc. 2006; 20(5):744–747. [PubMed: 16508817]
23. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL. Development of a model for training and evaluation of laparoscopic skills. Am J Surg. 1998; 175(6):482–487. [PubMed: 9645777]
24. Nakayama DK, Steiber A. Surgery interns' experience with surgical procedures as medical students. Am J Surg. 1990; 159(3):341–3. discussion 344. [PubMed: 2305945]
25. Klingensmith ME, Brunt LM. Focused surgical skills training for senior medical students and interns. Surg Clin North Am. 2010; 90(3):505–518. [PubMed: 20497823]
26. Brunt LM, Halpin VJ, Klingensmith ME, Tiemann D, Matthews BD, Spitler JA, Pierce RA. Accelerated skills preparation and assessment for senior medical students entering surgical internship. J Am Coll Surg. 2008; 206(5):897–904. discussion 904–7. [PubMed: 18471719]

27. Stewart RA, Hauge LS, Stewart RD, Rosen RL, Charnot-Katsikas A, Prinz RA. Association for Surgical Education. A CRASH course in procedural skills improves medical students' self-assessment of proficiency, confidence, and anxiety. Am J Surg. 2007; 193(6):771–773. [PubMed: 17512294]
28. Hu Y, Tiemann D, Michael Brunt L. Video self-assessment of basic suturing and knot tying skills by novice trainees. J Surg Educ. 2013; 70(2):279–283. [PubMed: 23427977]
29. MacDonald J, Williams RG, Rogers DA. Self-assessment in simulation-based surgical skills training. Am J Surg. 2003; 185(4):319–322. [PubMed: 12657382]
30. Brydges R, Carnahan H, Rose D, Dubrowski A. Comparing self-guided learning and educator-guided learning formats for simulation-based clinical training. J Adv Nurs. 2010; 66(8):1832–1844. [PubMed: 20557388]
31. Hildebrand C, Trowbridge E, Roach MA, Sullivan AG, Broman AT, Vogelman B. Resident self-assessment and self-reflection: University of Wisconsin-Madison's Five-Year Study. J Gen Intern Med. 2009; 24(3):361–365. [PubMed: 19156469]
32. Stewart J, O'Halloran C, Barton JR, Singleton SJ, Harrigan P, Spencer J. Clarifying the concepts of confidence and competence to produce appropriate self-evaluation measurement scales. Med Educ. 2000; 34(11):903–909. [PubMed: 11107014]
33. Gow KW. Self-evaluation: how well do surgery residents judge performance on a rotation? Am J Surg. 2013; 205(5):557–62. discussion 562. [PubMed: 23499389]
34. Xeroulis GJ, Park J, Moulton CA, Reznick RK, Leblanc V, Dubrowski A. Teaching suturing and knot-tying skills to medical students: a randomized controlled study comparing computer-based video instruction and (concurrent and summary) expert feedback. Surgery. 2007; 141(4):442–449. [PubMed: 17383520]
35. Mandel LS, Goff BA, Lentz GM. Self-assessment of resident surgical skills: is it feasible? Am J Obstet Gynecol. 2005; 193(5):1817–1822. [PubMed: 16260241]
36. Scott DJ, Goova MT, Tesfay ST. A cost-effective proficiency-based knot-tying and suturing curriculum for residency programs. J Surg Res. 2007; 141(1):7–15. [PubMed: 17574034]
37. Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. Am J Surg. 2006; 192(3):372–378. [PubMed: 16920433]
38. Brydges R, Carnahan H, Dubrowski A. Assessing suturing skills in a self-guided learning setting: absolute symmetry error. Adv Health Sci Educ Theory Pract. 2009; 14(5):685–695. [PubMed: 19132540]
39. Arnold L, Willoughby TL, Calkins EV. Self-evaluation in undergraduate medical education: a longitudinal perspective. J Med Educ. 1985; 60(1):21–28. [PubMed: 3965720]
40. Fitzgerald JT, White CB, Gruppen LD. A longitudinal study of self-assessment accuracy. Med Educ. 2003; 37(7):645–649. [PubMed: 12834423]
41. Ward M, MacRae H, Schlachta C, Mamazza J, Poulin E, Reznick R, Regehr G. Resident self-assessment of operative performance. Am J Surg. 2003; 185(6):521–524. [PubMed: 12781878]

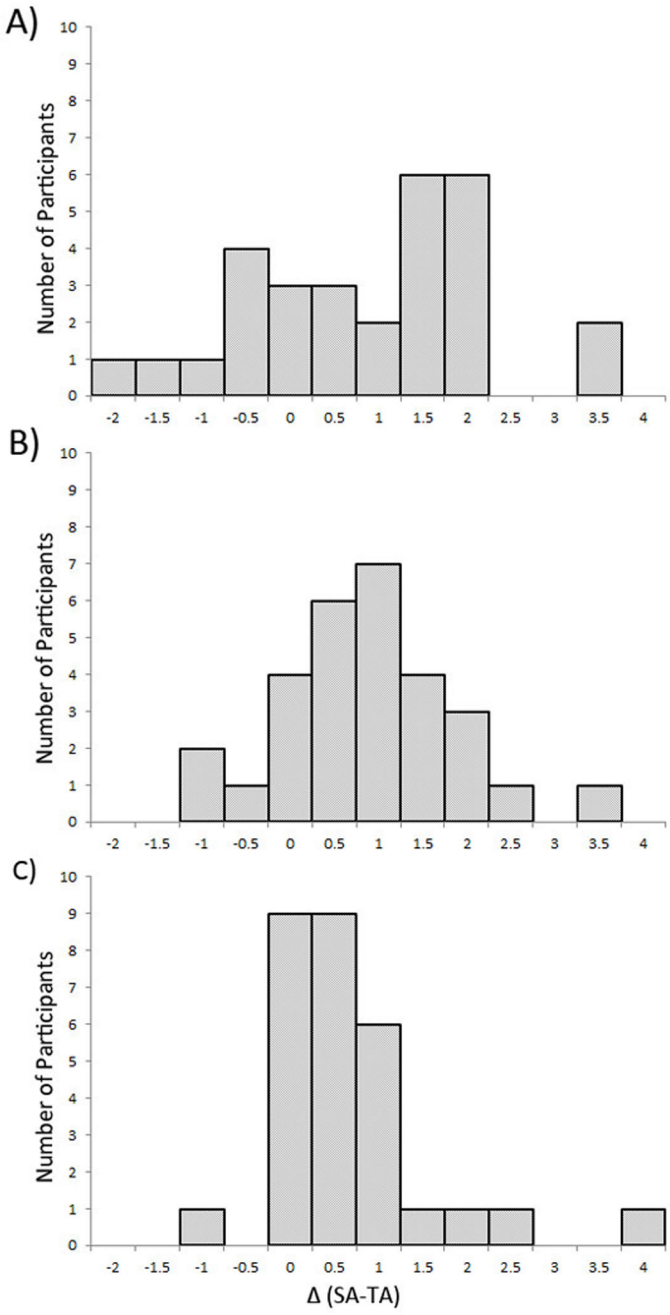

Figure 1.

Average difference between in experienced self- and instructor-assessments (SA-TA) for each participant, by tertiles of practice volume. With increasing experience, self-assessments trend from patterns of overestimation in the first tertile (A) to improved accuracy by the third tertile (C).

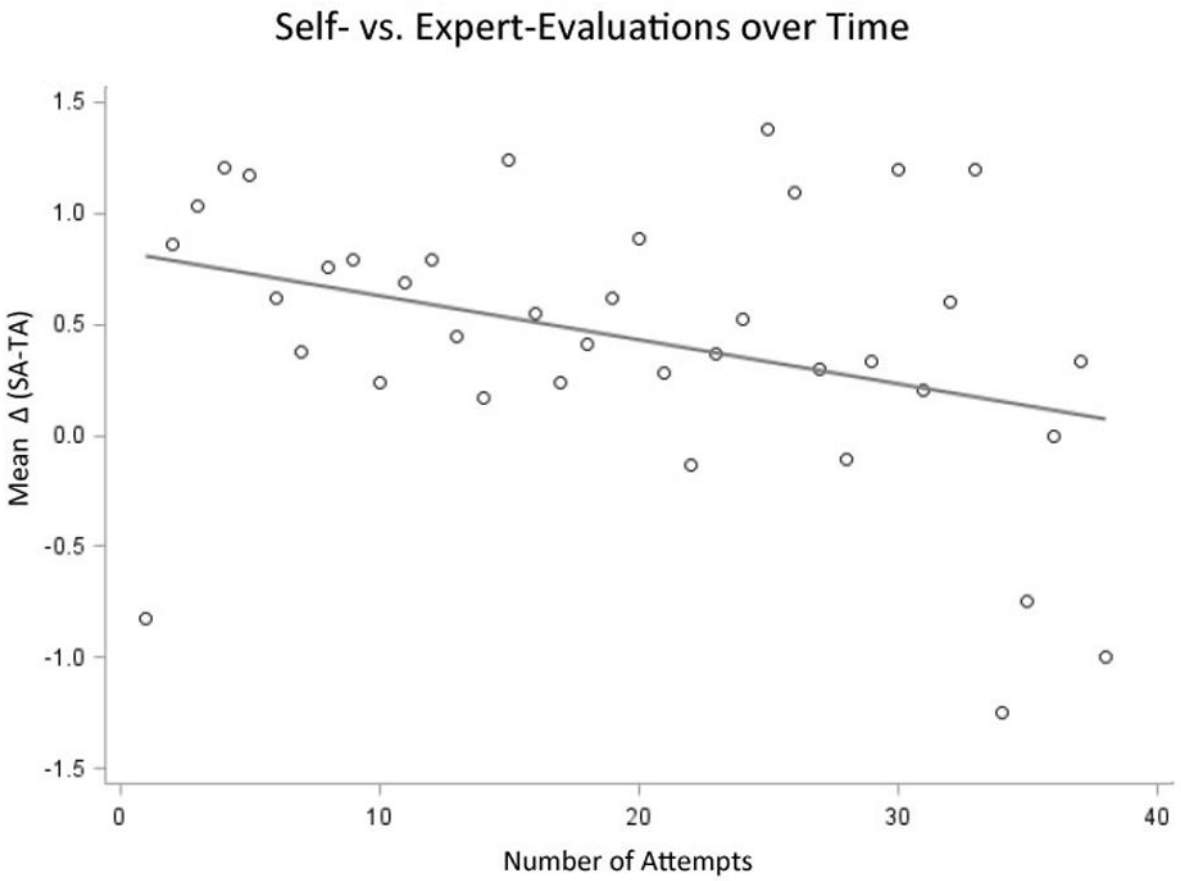

Figure 2.

Change in average difference (Δ) between self- and trainer-assessments (SA-TA) with increasing task attempts, across all participants. The positive difference between SA and TA decreases with greater practice experience ($p = 0.031$).

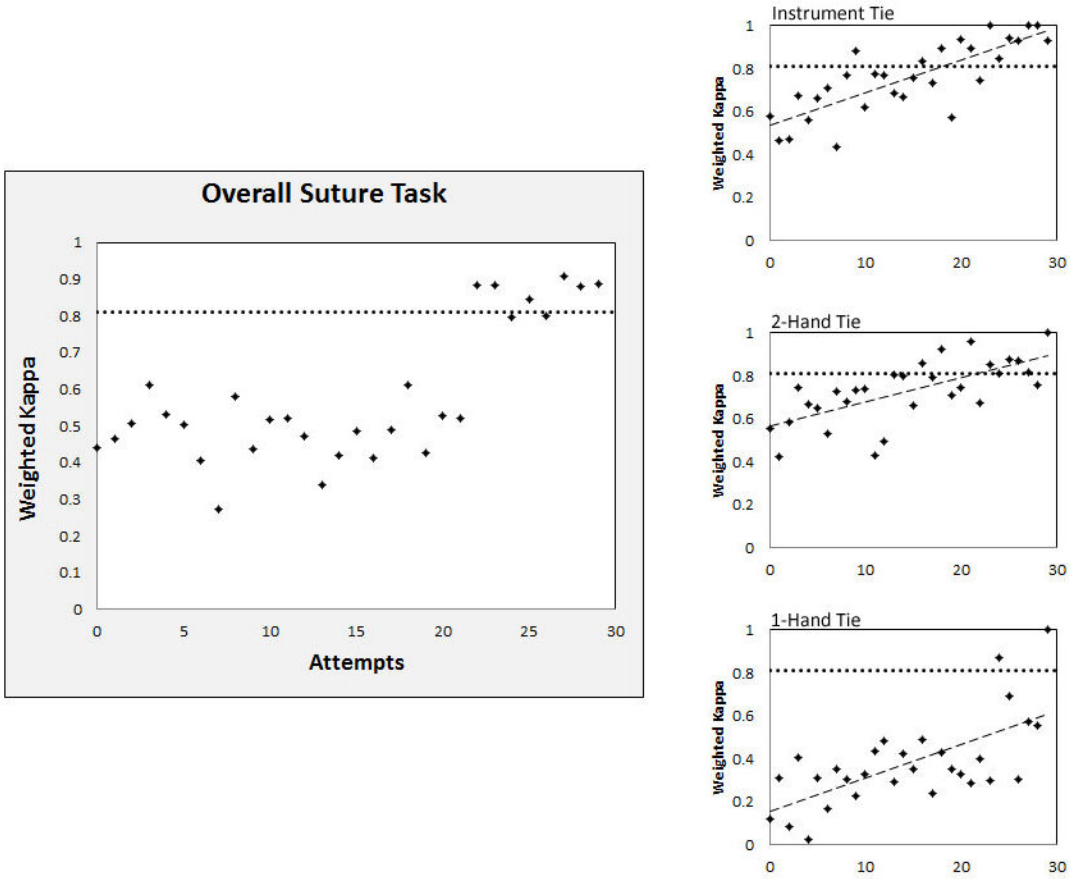

Figure 3.

Average agreement between self- and trainer-assessments with increasing task attempts across all participants, measured by Cohen's weighted kappa coefficient. Threshold for excellent agreement (weighted kappa ≥ 0.81, dotted line) is shown for comparison

Table 1

Technique scoring criteria used by participants, trainers, and faculty

Suturing Task – Technical Criteria	
Overall Technique (10)	
Does not grasp needle with hand	3
Loads needle 90°, past mid-arch	1
Turns needle through tissue	1
≤ 3 dots missed	3
Needle free when tying	1
Needle does not pass through knot	1
Two-handed Knot (9)	
First knot is surgeon's knot	1
Flat knots (crosses hands)	1
Alternates knots appropriately	1
Does not drop suture	1
No air knot	5
One-handed Knot (8)	
Pushes knots down with 1st finger	1
Alternates knots appropriately	1
Does not drop suture	1
No air knot	5
Instrument Tie (9)	
Tail length ≤2 cm	1
Grasps tail at distal 1/2 cm	1
First knot is surgeon's knot	1
Flat knots (crosses hands)	1
No air knot	5
Maximum Score	**36**

Table 2

Subgroup analysis of differences (Δ) between self- and trainer-assessments by skill proficiency.

Initial Proficiency[a]	High (N=16)	Low (N=13)	p-value
ΔComposite	0.11	1.34	0.01
Δ2-Handed	−0.09	0.22	0.25
Δ1-Handed	0.03	0.47	0.04
ΔInstrument	0.05	0.34	0.30

Post-Test Proficiency[b]	High (N=13)	Low (N=16)	p-value
ΔComposite	0.48	0.81	0.52
Δ2-Handed	0.01	0.07	0.81
Δ1-Handed	0.28	0.18	0.62
ΔInstrument	0.04	0.30	0.36

[a] Average of first three trainer assessments, relative to median

[b] Post-test evaluation score, relative to median