The American Journal of Surgery 220 (2020) 610–615

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

The impact of rater training on the psychometric properties of standardized surgical skill assessment tools

Reagan L. Robertson, Jason Park, Lawrence Gillman, Ashley Vergis*

Department of Surgery, University of Manitoba, Canada

ARTICLE INFO

Article history:
Received 6 May 2019
Received in revised form
9 January 2020
Accepted 10 January 2020

Keywords:
Assessment
Surgical education
Rater training
Technical skills

ABSTRACT

Introduction: Competency-based frameworks are common in surgical training. However, the optimal use of standardized technical assessments is not well defined. We investigated the effect of rater training (RT) on the reliability and validity of four assessment tools.

Materials and methods: Forty-Seven surgeons were randomized to RT (N = 24) and no training (N = 23) groups. A task-specific checklist, pass-fail, visual analog, and OSATS global rating scale (GRS) were used to assess trainee knot-tying and suturing tasks. Delayed assessment was performed two weeks later. Internal consistency, intra/inter-rater reliability, and construct validity were measured.

Results: The GRS had superior reliability and validity compared to the other tools regardless of training. No significant differences between training groups was found. However, the RT group trended to improved reliability for all tools at both assessments.

Conclusions: RT did not lead to significant improvements in skills assessments. Standardized assessments (OSATS GRS) are preferred due to their superior reliability and validity over other methods. Despite findings, we believe more effective training methods or repeated sessions may be required for sustained and significant effects.

Introduction

Rater Training (RT) was developed to address the natural bias introduced by subjective performance assessments, as reliable and valid assessment methods are critical to ensure fair and accurate evaluation and advancement. This is particularly true as training programs shift to competency based training models.[1,2] RT is a process whereby raters undergo instruction on how to best evaluate trainees and produce reliable and accurate scores.[3] This is paramount in surgical training as robust technical assessment tools are needed to assess residents' ability to operate independently and complete training. Unfortunately, most competency frameworks do not specifically assess technical competency and these skills are somewhat arbitrarily grouped under categories such as "medical expert" or "patient care".[4] However, technical competence is clearly too essential to be evaluated within another category for surgical disciplines. Well-designed and high-quality methods of assessment specific to technical skills are thus required.

More objective and structured assessments have been designed given this need. These include visual analog scales (VAS), task-specific checklists, global rating scales (GRS) and combined assessments.[5–10] Perhaps the most well studied of these tools is the Objective Structured Assessment of Technical Skill (OSATS) which is considered by many to be the gold standard of technical skill evaluation.[11] However, many of these tools are suboptimal in terms of their psychometric properties and levels of reliability and validity.[11]

RT is one method that attempts to improve rater objectivity by systematically improving the reliability of assessments.[12] However, while RT improves the reliability and validity of observational assessment tools,[3] it appears to have a more variable effect in medical education.[13–16] Despite this, many studies continue to highlight the importance of RT prior to the use of technical skill assessment tools.[11,17] This study assessed the effect of RT on the reliability and validity of four established technical skill assessment tools.

* Corresponding author. St. Boniface General Hospital, Z-3039 - 409 Tache Avenue, Winnipeg, MB, Canada, R2H 2A6.
E-mail address: avergis@sbgh.mb.ca (A. Vergis).

https://doi.org/10.1016/j.amjsurg.2020.01.019
0002-9610/

Methods

Study design

The University of Manitoba Health Research Ethics Board approved this study. Surgeons with certification in any Royal College of Physician and Surgeons of Canada surgical subspecialty were eligible for voluntary participation. Raters evaluated videos of trainees performing a suturing and knot tying task. Participants were randomized to RT and to a no training control group using stratified block randomization, controlling for subspecialty and practice setting. The RT group underwent frame-of-reference training, which has been shown to be the most effective method of training in previous reviews.[3] Participants in the control group proceeded directly to trainee evaluation. A second rating session was performed with all raters approximately two weeks after the initial session to determine if there was a sustained effect of training. Neither group underwent any form of training prior to the second session. After the second rating session, raters in the control group were offered the opportunity to watch the training video if desired.

Rater training intervention

A brief frame-of-reference training video was developed, and has been previously published (https://youtu.be/CzF-hEywufQ).[18] Frame-of-reference training instructs raters on performance standards for an assessment tool. The desired level of performance for each rating was explained to create a shared definition between raters of an appropriate ranking for an observed performance. Three surgeons with graduate degree training in medical education reviewed the video prior to use for training.

Trainee assessments

The task chosen for evaluation was simple suture and instrument tie. This basic and widely applicable task was chosen to facilitate recruitment and training of a variety of surgeons. It also allowed for evaluation of multiple trainees in a single session. Ten videos of trainees performing a simple suture and instrument tie on a plastic model were developed, including a range of training levels from third year medical students to third year general surgery residents. Only the trainee's gloved hands and operative field were shown to allow for blinded assessments. Videos were shown to each rater in a random sequence at both assessments. Raters watched each video one full time through and then completed the assessment tools.

Assessment tools

Raters used four assessment tools to evaluate trainees' technical skills: (1) a pass-fail designation, (2) a VAS, (3) a task-specific checklist, and (4) a GRS modified from OSATS. The data collection sheet has been previously published.[18] The pass-fail designation required a dichotomous rating of either pass or fail for the overall trainee performance. The VAS used a 10 cm horizontal line with verbal descriptors at each end. The rater places a mark on the line to indicate the level of overall technical skill (maximum possible score 10.0). The task-specific checklist consisted of a list of 10 steps required to correctly perform a suturing and knot-tying task. One point is assigned for each item performed correctly and zero points are given for incorrect items or items not performed (maximum possible score 10).[9,10] Finally, the OSATS GRS was modified, as certain aspects of the original scale could not be evaluated on the trainee assessment videos, including "knowledge of instruments" and "use of assistants". The final adapted scale assessed respect for tissue, time and motion, instrument handling, flow of procedure and overall performance.[7,19] Each items is rated on a 5-point scale with verbal descriptors at points 1, 3, and 5. The scores from each item were averaged to give the final score (maximum possible score 5.0).

Statistical analysis

Analyses were completed using SPSS (Version 17.0) and R Console (Version 3.1.0). Internal consistency for the multi-item scales (checklist and GRS) was measured with Cronbach's alpha. Reliability was measured with Fleiss Kappa coefficient for the pass-fail designation. For the continuous variables, inter-rater reliability (IRR). Initial and delayed rater agreement were calculated using intraclass correlation (ICC) type 2. Construct validity was assessed using univariate logistic regression for each assessment tool score relative to senior compared to junior trainee level. An interaction term between the variable and RT group was included in the regression to determine if there was any effect of RT on construct validity. The p-value of the interaction term in the model determined if a significant effect of RT was present. Multivariate analysis was then performed for "senior" trainee level using all assessment tool scores concurrently for each RT group controlling for rater characteristics.

Results

Forty-seven surgeons were randomized to RT and control groups. Surgeons were recruited from a variety of sub-specialties including General Surgery, Urology, Orthopedics, Otolaryngology, Neurosurgery, Thoracic, Cardiac and Plastic Surgery. As previously published, there was no significant difference between training groups in terms of surgeon specialty, practice setting, gender, age, experience with the training tools, and number of trainees per year.[18] There was a small difference in years in practice between the no training group (9.5± 7.3) and Rater Training group 14.4± 8.2 ($p = 0.04$). No surgeons had previous experience with RT.

Internal consistency

Internal consistency for the GRS was excellent for both training groups at the initial and delayed assessments and inter-item correlations were good to excellent. Checklist items were associated with much lower levels of internal consistency for both training groups, and correlations were distributed over a wide range (Table 1). Further analysis did not support deleting any of the individual checklist items, as deleting any one item did not lead to improvement in Cronbach's alpha for the overall scale.

Reliability

Initial and delayed rater agreement was acceptable for the GRS and VAS assessment forms and low for the checklist and pass-fail designation. There was a trend towards higher agreement for the RT group but this was not statistically significant (Table 2).

A statistically significant improvement in IRR was not demonstrated for any of the four assessment forms. IRR was uniformly poor for the pass-fail designation (Table 3). IRR for the continuous assessment tools at the initial assessment has been previously reported (VAS: RT = 0.71 (0.50–0.91) vs. No training (NT) = 0.46 (0.27–0.75); Checklist: RT = 0.46 (0.27–0.75) vs. NT = 0.33 (0.17–0.64); GRS: RT = 0.71 (0.52–0.89) vs. NT = 0.61 (0.41–0.85) (IRR with 95% CI)).[18] IRR for the continuous tools at the delayed assessment are shown in Table 3. Reliability was higher for the RT

Table 1
Internal consistency for multi-item assessment tools.

	No training		Rater Training	
	Cronbach's alpha	Mean Inter-item correlation and range	Cronbach's alpha	Mean Inter-item correlation and range
Checklist				
Initial	0.64	0.15 (−0.07-0.64)	0.64	0.16 (−0.06-0.54)
Delayed	0.63	0.15 (−0.02-0.7)	0.65	0.16 (−0.09-0.50)
GRS				
Initial	0.94	0.74 (0.56-0.85)	0.96	0.82 (0.73-0.88)
Delayed	0.93	0.73 (0.56-0.88)	0.95	0.80 (0.70-0.87)

group on all three continuous assessment tools at both assessments; however the 95% confidence intervals were wide and overlapped. There was a trend towards reliability being higher at the initial assessment, which was also not significant.

Validity

All assessment tools showed evidence of construct validity on univariate analysis, with higher scores being associated with increased odds of senior level of training (Table 4). Trained raters had a trend towards improved validity, with higher OR for training level. However, this was not statistically significant for any of the tools. On multivariate regression, only the GRS retained significant construct validity for both the training and control groups (Table 5). In the control group, improved checklist scores were significantly associated with junior level of training on multivariate analysis.

Discussion

The need to develop standardized high-quality tools to measure technical skill in surgical training is clear but doing so remains challenging. Assessment tools must capture multiple complex aspects of technical skill such as dexterity, judgment and knowledge over a range of procedures.[20] A recent review found that minimum performance levels to determine competence are often completely arbitrary across a heterogeneous mix of tools used for skill assessments.[11] Although OSATS represents the current "gold standard" of technical skill assessment, it fundamentally remains an observational tool despite the structured format. Even after extensive validation, OSATS at times fails to achieve the minimum accepted reliability of 0.8 for high stakes testing.[18,21,22] Efforts to improve standardized tools and define benchmarks for success remains of utmost importance as surgical education moves towards competency-based assessment.[23,24] Also of importance is that implementation of these tools into a training curriculum requires significant resources. It may be difficult for programs to know which tools to rely on especially given the number of different potential assessment mechanisms.

This study sought to examine the psychometric properties of several established technical assessment tools, and the effect of RT on these properties using a randomized, controlled design. Despite trends towards improved reliability and validity with RT, a significant difference between groups was not demonstrated. It may be difficult to adequately power reliability studies with multiple raters to show such differences.[18] Regardless, the results reveal important information about the psychometric properties of these rating tools which can guide their future use and study. Fig. 1 summarizes the psychometric properties of pass-fail ratings, VAS, checklists, GRS, and the effects of RT and timing of assessment on these tools.

Pass-fail designation

Use of the pass-fail designation was very heterogeneous, supporting the inconsistency of pass-fail assessments. Rater agreement and reliability was uniformly poor, consistent with prior study.[25] Results of the pass-fail designation do however provide insight into rater behavior. Particular raters passed every learner, regardless of their training group or the trainees' concurrent assessment tool scores. The same raters passed all trainees at both assessments in the majority of cases. Additionally, nearly a third of raters in both groups gave a passing grade to a learner with a clearly insufficient performance of the task. Such behavior could be used to identify raters who are excessively lenient and prone to rating errors, or those that are resistant to training. This suggests that pass-fail designations alone should not be used to make judgments about technical skill due to their unreliable and subjective nature.

Task-specific checklists

Checklist items had moderate internal consistency with poorly related and wide-ranging inter-item correlations. It has been suggested that checklists punish efficiency and innovation by giving points for proceeding methodically as opposed to measuring true competence.[26] Competent, high-level trainees who bypass the step-wise construct of the checklist may in fact then have lower checklist scores.[9,26] This characteristic may also convert experienced raters into observers that record binary events as the subtleties of experienced judgment are removed from the evaluation.[27] These phenomena may explain the variation in checklist scores, including why higher scores were not associated with a higher level of training on multivariate regression. As well, it may have been challenging for raters to accurately recall and capture each item for

Table 2
Assessment tool initial and delayed rater agreement.

	No training	Rater Training
Pass-fail	0.41 (0.26-0.57)	0.45 (0.32-0.59)
VAS	0.62 (0.54-0.70)	0.71 (0.64-0.77)
Checklist	0.46 (0.35-0.56)	0.53 (0.43-0.61)
GRS	0.66 (0.58-0.73)	0.73 (0.67-0.79)

Table 3
Inter-rater reliability of assessment tools (ICC).

Assessment Form	IRR with 95% confidence interval	
	No training	Rater Training
Pass-fail		
Initial	0.20 (0.17-0.25)	0.22 (0.18-0.26)
Delayed	0.11 (0.07-0.15)	0.16 (0.13-0.20)
VAS		
Delayed	0.43 (0.24-0.73)	0.54 (0.34-0.80)
Checklist		
Delayed	0.27 (0.13-0.57)	0.42 (0.23-0.71)
GRS		
Delayed	0.52 (0.32-0.79)	0.64 (0.45-0.86)

Table 4
Univariate logistic regression odds ratios for senior level of training.

Assessment tool	No Training	Rater Training	p-value
Pass/Fail	5.39 (2.53-11.46)	6.08 (3.08-12.01)	0.82
VAS	1.45 (1.27-1.65)	1.70 (1.46-1.97)	0.12
CT	1.23 (1.07-1.41)	1.50 (1.30-1.73)	0.06
GRS	3.73 (2.51-5.54)	4.25 (2.85-6.33)	0.65

a relatively brief task. Although accuracy was not specifically addressed in this study, one trainee failed to cut the sutures and yet close to half of raters in both training groups marked this task as complete on the checklist. All of these factors had the potential to limit the reliability of the checklist. IRR and initial and delayed rater agreement were fair to moderate, performing only minimally better than the pass-fail designation. This raises concern about the use of checklists as a sole means of trainee evaluation. However, checklists still remain one of the limited ways to measure procedure-specific knowledge in a systematic fashion. This suggests that combining a task-specific checklist with a second, more reliable overall assessment tool should be considered if evaluation of procedure-specific information is required.

VAS

VAS had IRR and initial and delayed rater agreement most comparable to the GRS in the current study. VAS were also associated with the largest difference in IRR between the trained and untrained groups. These tools are intuitive to use, but by design are relatively simple with minimal additional descriptors provided. This may cause VAS to be more susceptible to different interpretations of the upper and lower ends of the scale by individual raters. Defining the upper and lower limits with anchors may have helped improve the shared understanding of these limits and increased reliability to levels approaching the GRS. VAS's function similarly to GRS in that they measure a perceived general concept of technical skill. They are also intrinsically easy to use and understand. Consequently, they may not add any additional information to what is already evaluated with a GRS. When assessed in multivariate analysis with the other tools, the VAS did not maintain evidence of significant construct validity. These tools may then serve as a fast and uncomplicated assessment of technical skill. However, they may not outperform other well-established tools such as the GRS.

GRS

Internal consistency was excellent for the OSATS GRS with values similar to previously published in the literature.[19,22,25] One of the innate advantages of GRS is that the tool measures several related qualities that represent an overall construct of technical ability. There were trends towards improved reliability on the GRS with RT. The GRS specifies that it is to be used "irrespective of training level". This was emphasized throughout the training. We sought to more clearly outline performance standards for each portion of the GRS. The upper limit of the scale was defined as the performance one would expect from a competent surgeon as that is the ultimate goal of training. This definition is not without flaw though as variability within practicing competent surgeons is still expected. However, this portion of the definition was meant to emphasize the correct reference for the scale's upper limit. Surgeons not explicitly trained in the optimal use of the assessment tools may erroneously use the top end of the scale as the best performance they would expect from an individual trainee or of trainees at a particular training level. The GRS contains anchoring statements but in many cases they include terms not inherently understood or representative of a specific level of performance. Although this characteristic may decrease reliability between users, it also allows the GRS to be applied to many different procedures and settings. This makes some degree of training essential to create procedure-specific shared definitions between raters and ensure reliable use in varied situations.[11,28] The GRS had the highest IRR of all tools in both training groups at each assessment. Initial and delayed rater agreement was good for both groups. GRS were also the only evaluation tool to maintain evidence of construct validity in the multivariate regression analysis. Therefore, this study supports the body of evidence advocating the use of the OSATS GRS as the current optimal technical skill assessment tool.[11]

Table 5
Multivariate logistic regression odds ratios for senior level of training.

Assessment tool	Odds Ratio (95% confidence interval)	p-value
No Training		
Pass-Fail	0.46 (0.16-1.34)	0.15
VAS	1.02 (0.78-1.35)	0.86
Checklist	0.64 (0.50-0.82)	<0.001
GRS	6.31 (3.05–13.05	<0.001
Rater Training		
Pass-Fail	0.88 (0.34-2.28)	0.79
VAS	1.15 (0.87-1.51)	0.33
Checklist	0.94 (0.76-1.17)	0.59
GRS	3.31 (1.60-6.84)	<0.001

Limitations

There are several limitations in this study which may have impacted the psychometric properties of the assessment tools. Reliability may improve with a broader range of observed performances.[12] Evaluation of a more complex task or a wider range of training levels thus may have increased reliability. Methodological challenges occurred during completion of some of the tools. It was difficult to complete all 4 tools in a timely and accurate manner for the simple surgical task, in particular longer tools such as the checklist. Three surgical education experts provided feedback on the RT protocol, but did not perform trial ratings. This addition may have identified some of the challenges raters faced when using the tools and allowed for improvements. Use of a modified version of the OSATS GRS was necessary given that some of the factors could not be determined from our videos. While this practice is common in surgical education studies, it necessitates re-validate of the tool and limits the generalizability of results and the ability to compare results to other studies that use the full scale or different modifications.[11] As longer tools generally have superior reliability, removing points from the GRS could also have contributed to decreased reliability.

Developing an effective RT protocol is challenging due to variability in the existing literature about the optimal training format.[18] We sought to develop a concise training video that could be shown prior to any individual training session. However, the training format of the present study may have been sub-optimal or too brief to lead to significant changes in rater behavior.[18] Longer or multiple training sessions tools may have been necessary to integrate the training construct. Training for a smaller number of tools may have been more effective. Practice rating sessions with feedback or video examples of trainee performance levels could also have been of benefit. Additionally, RT could have been repeated at the delayed assessment to see if there was any additive or sustained effect, as opposed to assessing if a potential effect was lost over time. It is impossible to know without further study if changes to the current training would lead to additional improvement in the psychometric

Result	Assessment Tool	Summary of Findings
Internal Consistency		
	Pass-Fail	N/A
	VAS	N/A
	Checklist	Good agreement No change with RT, no change with assessment time
	GRS	Excellent agreement No change with RT, no change with assessment time
Rater Agreement		
	Pass-Fail	Moderate agreement No change with RT
	VAS	Good agreement Trend to improvement with RT
	Checklist	Moderate agreement Trend to improvement with RT
	GRS	Good agreement Trend to improvement with RT
IRR		
	Pass-fail	Poor agreement Trend to improvement with RT Trend to higher scores at initial assessment
	VAS	Moderate agreement Trend to good agreement with RT Trent to higher scores at initial assessment
	Checklist	Fair agreement Trend to moderate agreement with RT Trend to higher scores at initial assessment
	GRS	Good agreement Trend to improvement with RT Trend to higher scores at initial assessment
Validity		
	Pass-fail	Significant validity on univariate analysis Not maintained on multivariate analysis Trends to improvement with RT
	VAS	Significant validity on univariate analysis Not maintained on multivariate analysis Trends to improvement with RT
	Checklist	Significant validity on univariate analysis Higher scores associated with junior training level on multivariate analysis in the no training group
	GRS	Significant validity on univariate analysis Significant validity on multivariate analysis Trends to improvement with RT

Excellent agreement 0.81-1.0, Good agreement 0.61-0.8, Moderate agreement 0.41-0.6, fair agreement 0.21-0.4, poor agreement <0.2.
Adapted from Gwet K L. *Handbook of Inter-Rater Reliability*. 3rd Editio. Gairthersburg, MD: Advanced Analytics, LLC; 2012.
RT – rater training; VAS – visual analog scale; GRS – global rating scale; IRR – inter-rater reliability

Fig. 1. Summary of assessment tool reliability and validity results.

properties of the tools, or if an entirely new training format would be beneficial.

Finally, our group of untrained raters had been in practice slightly longer than the RT group, although there was a wide range in both groups. Regardless, experience with assessment tools and numbers of trainees were similar between groups, which likely represent more important factors for determining whether one is an "expert" evaluator.

Despite these limitations, the authors believe that this investigation is one of the most robust examples of rater training in surgical education and that these results provide important insight into psychometric properties of standardized technical skill assessments. It is likely that many clinically active surgeons and training programs underutilize these assessment methods due to perceived constraints such as lack of tool familiarity or the need for formal training. The results of this study suggest that experienced surgeons can evaluate performance with reasonable accuracy based on their experience. Thus, it is important that educators and programs are encouraged to use standardized skills assessments like the GRS in surgical training while efforts to refine and improve them are on-going.

Conclusions

This paper highlights some of the limitations of current technical skills assessments. Rater training was not associated with a significant improvement in the psychometric properties of common technical skill assessment tools. The reliabilities of pass-fail designations and task-specific checklists were poor, and results from these tools should be interpreted with caution. For the VAS and GRS, the reliabilities demonstrated would be considered "good" for the RT group and "moderate" for the non-trained group for educational purposes. Despite our efforts to improve reliability, the IRR of all the tools remained below the minimum threshold of 0.8 required for high stakes testing. Although all tools showed evidence of validity by construct, this was only maintained by the GRS on multivariate analysis. Our findings support the continued use of the GRS as the preferred method of technical skill evaluation in surgical training given its superior reliability and validity.

Contributions

Conception and design, or data acquisition, or data analysis and interpretation: Robertson, Vergis, Gillman, Park; Drafting or critical revisions: Robertson, Vergis, Gillman, Park:,Final approval: Robertson, Vergis, Gillman, Park.

Declaration of competing interest

The authors have no conflicts of interest to disclose.

References

1. Gray JD. Global rating scales in residency education. *Acad Med*. 1996;71(1 Suppl):S55–S63. https://doi.org/10.1097/00001888-199601000-00043.
2. Reznick RK. Teaching and testing technical skills. *Am J Surg*. 1993;165(3):358–361. https://doi.org/10.1016/S0002-9610(05)80843-8.
3. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occup Organ Psychol*. 1994;67:189–205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x.
4. Frank JR, Danoff D, Frank JR, Danoff D. *The CanMEDS Initiative : Implementing an Outcomes-Based Framework of Physician Competencies the CanMEDS Initiative : Implementing an Outcomes-Based Framework of Physician Competencies*. 2015. https://doi.org/10.1080/01421590701746983. October.
5. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107–113. https://doi.org/10.1016/j.amjsurg.2005.04.004.
6. Vassiliou MC, Kaneva PA, Poulose BK, Dunkin BJ. *Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): A Valid Measurement Tool for Technical Skills in Flexible Endoscopy*. 2010:1834–1841. https://doi.org/10.1007/s00464-010-0882-8.
7. 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. https://doi.org/10.1002/bjs.1800840237.
8. Huang E, Vaughn CJ, Chern H, Sullivan PO, Kim E. An objective assessment tool for basic surgical knot-tying skills. *J Surg Educ*. 2015;72(4):572–576. https://doi.org/10.1016/j.jsurg.2015.01.002.
9. Figert PL, Park AE, Witzke DB, Schwartz RW. Transfer of training in acquiring laparoscopic skills. *J Am Coll Surg*. 2001;193(5):533–537. https://doi.org/10.1016/S1072-7515(01)01069-9.
10. Rosser JC, Rosser LE, Savalgi RS. Skill acquisition and assessment for laparoscopic surgery. *Arch Surg*. 1997;132(2):200–204. https://doi.org/10.1001/archsurg.1997.01430260098021.
11. Szasz P, Louridas M, Harris K a, Aggarwal R, Grantcharov TP. Assessing technical competence in surgical trainees: a systematic review. *Ann Surg*; 2014, 00(00) http://www.ncbi.nlm.nih.gov/pubmed/25119118.
12. Wanzel KR, Ward M, Reznick RK. Teaching the surgical craft: from selection to certification. *Curr Probl Surg*. 2002;39(6):583–659. https://doi.org/10.1067/mog.2002.123481.
13. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140:874–881. https://doi.org/10.7326/0003-4819-140-11-200406010-00008 [pii].
14. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med*. 1992;117(9):757–765.
15. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. *Med Educ*. 1980;14(5):345–349. https://doi.org/10.1111/j.1365-2923.1980.tb02379.x.
16. George BC, Teitelbaum EN, DaRosa DA, et al. Duration of faculty training needed to ensure reliable or performance ratings. *J Surg Educ*. 2013;70:703–708. https://doi.org/10.1016/j.jsurg.2013.06.015.
17. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Continuing Educ Health Prof*. 2012;32(4):279–286. https://doi.org/10.1002/chp.21156.
18. Robertson R, Vergis A, Gillman L, Park J. Effect of rater training on the reliability of technical skill assessments: a randomized controlled trial. *Can J Surg*. 2018;61(6).
19. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173(3):226–230. https://doi.org/10.1016/S0002-9610(97)89597-9.
20. Darzi A, Mackay S. Assessment of surgical competence. 2001;10(Suppl II):64–70.
21. Sidhu R, Grober E, Musselman L, Reznick R. Assessing competency in surgery: where to begin? *Surgery*. 2004;135(1):6–20. https://doi.org/10.1016/S0039-6060(03)00154-5.
22. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73(9):993–997. https://doi.org/10.1097/00001888-199809000-00020.
23. The Royal College of Physicians and Surgeons of Canada. About competence by design (CBD). http://www.royalcollege.ca/rcsite/cbd/competence-by-design-cbd-e.
24. Holmboe ES, Sherbino J, Long DM, et al. *The Role of Assessment in Competency-Based Medical Education the Role of Assessment in Competency-Based Medical Education*. 2015. https://doi.org/10.3109/0142159X.2010.500704. October.
25. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skills (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
26. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. *Acad Med*. 1999;74(10):1129–1134. https://doi.org/10.1097/00001888-199910000-00017.
27. Reznick R, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D. High stakes examinations: what do we know about measurement? *Acad Med*. 1998;73(10):97–99.
28. Gallagher AG, Ritter EM, Satava RM. *Fundamental Principles of Validation , and Reliability : Rigorous Science for the Assessment of Surgical Education and Training*. 2003:1525–1529. https://doi.org/10.1007/s00464-003-0035-4.