Effect of rater training on the reliability of technical skill assessments: a randomized controlled trial

Reagan L. Robertson, MD, MSc
Ashley Vergis, MD, MMedEd
Lawrence M. Gillman, MD, MMedEd
Jason Park, MD, MEd

Abstract presented at the 36th Annual Meeting of the Association for Surgical Education, Boston, Apr. 12–14, 2016.

Accepted Jan. 30, 2018; Early-released Oct. 1, 2018; subject to revision

Correspondence to:
J. Park
St. Boniface General Hospital
Z-3031 – 409 Taché Ave
Winnipeg MB R2H 2A6
jpark@sbgh.mb.ca

DOI: 10.1503/cjs.015917

Background: Rater training improves the reliability of observational assessment tools but has not been well studied for technical skills. This study assessed whether rater training could improve the reliability of technical skill assessment.

Methods: Academic and community surgeons in Royal College of Physicians and Surgeons of Canada surgical subspecialties were randomly allocated to either rater training (7-minute video incorporating frame-of-reference training elements) or no training. Participants then assessed trainees performing a suturing and knot-tying task using 3 assessment tools: a visual analogue scale, a task-specific checklist and a modified version of the Objective Structured Assessment of Technical Skill global rating scale (GRS). We measured interrater reliability (IRR) using intraclass correlation type 2.

Results: There were 24 surgeons in the training group and 23 in the no-training group. Mean assessment tool scores were not significantly different between the 2 groups. The training group had higher IRR than the no-training group on the visual analogue scale (0.71 v. 0.46), task-specific checklist (0.46 v. 0.33) and GRS (0.71 v. 0.61). However, confidence intervals were wide and overlapping for all 3 tools.

Conclusion: For education purposes, the reliability of the visual analogue scale and GRS would be considered "good" for the training group but "moderate" for the no-training group. However, a significant difference in IRR was not shown, and reliability remained below the desired level of 0.8 for high-stakes testing. Training did not significantly improve assessment tool reliability. Although rater training may represent a way to improve reliability, further study is needed to determine effective training methods.

Contexte : La formation des évaluateurs améliore la fiabilité des outils d'évaluation observationnels, mais n'a pas été rigoureusement étudiée au plan des habiletés techniques. Cette étude a tenté de vérifier si la formation des évaluateurs permettait d'améliorer la fiabilité de l'évaluation des habiletés techniques.

Méthodes : On a assigné des chirurgiens universitaires et communautaires appartenant aux surspécialités chirurgicales du Collège royal des médecins et chirurgiens du Canada, soit à une formation des évaluateurs (vidéo de 7 minutes comprenant des éléments de formation afférents au cadre de référence), soit à l'absence de formation. les participants ont ensuite évalué des stagiaires qui effectuaient tâches, telles sutures et nœuds, à l'aide de trois outils d'évaluation : échelle analogique visuelle, liste de vérification spécifique à la tâche et version modifiée de l'échelle d'appréciation globale (ÉAG) de l'Objective Structured Assessment of Technical Skill. Nous avons mesuré la fiabilité interévaluateurs (FIÉ) à l'aide de la corrélation intraclasse de type 2.

Résultats : Il y avait 24 chirurgiens dans le groupe soumis à la formation et 23 dans le groupe non soumis à la formation. Les scores moyens des outils d'évaluation n'ont pas été significativement différents entre les deux groupes. Le groupe soumis à la formation a présenté une FIÉ plus élevée que l'autre groupe à l'échelle analogique visuelle (0,71 c. 0,46), à la liste de vérification spécifique à la tâche (0,46 c. 0,33) et à l'ÉAG (0,71 c. 0,61). Par contre, les intervalles de confiance étaient larges et se recoupaient pour les trois outils.

Conclusion : Aux fins de la formation, la fiabilité de l'échelle analogique visuelle et de l'ÉAG serait considérée « bonne » pour le groupe soumis à la formation, mais « modérée » pour le groupe non soumis à la formation. On n'a toutefois pas démontré de différence significative quant à la FIÉ et la fiabilité est demeurée inférieure au niveau souhaité de 0,8 pour les tests importants. La formation n'a pas significativement amélioré la fiabilité de l'outil d'évaluation. Même si la formation des évaluateurs représente potentiellement une façon d'améliorer la fiabilité, il faudra approfondir la recherche pour déterminer quelles méthodes de formation sont efficaces.

High-quality assessments of technical skill form an essential component of trainee evaluation for surgical training programs. Technical skill assessments help educators follow trainee progress, identify training deficiencies and determine the effects of teaching interventions. Despite their obvious importance and complexity, technical skill assessments in many training programs have long remained subjective, unstandardized and informal.[1] Traditionally, programs relied on subjective faculty assessments provided over the duration of training to determine competence. This method lacks rigorous criteria, is prone to systematic bias and may not correlate to other forms of assessment.[2–4] Recognizing the limitations of surgical skill assessments, some groups have developed standardized assessment tools to further improve technical evaluations.[3,5–7] Notably, Reznick and colleagues[5] at the University of Toronto developed the Objective Structured Assessment of Technical Skill (OSATS), which uses a 5-point Likert global rating scale (GRS) with behavioural anchors to measure 7 aspects of technical skill. The OSATS GRS is one of the most widely studied and validated surgical skill assessment tools and is viewed by many as the current gold standard for technical skill evaluation.[2,8]

Despite the widespread use and popularity of the OSATS, the tool has notable limitations that reduce its reliability. The OSATS remains fundamentally observational, with an inherent subjectivity in ratings despite its structured format. It is vulnerable to common rating errors, which raters may be more prone to when not trained in the correct use of the tool.[4] There are multiple factors that may contribute to decreasing interrater reliability (IRR) when 2 or more raters observe and score a performance. Previous interactions with trainees may cause either positive or negative bias. Raters may also have different interpretations of terms on the tool or give different weight to certain aspects of the procedure. Finally, raters may use disparate criteria to judge performance or have different standards for performance relative to training level. The current published IRR for the OSATS is 0.64–0.72.[3,9] These values consistently fall below the optimal value of 0.8 desired for high-stakes testing.[10,11] Therefore, room for improved reliability remains, even for one of the most extensively validated surgical skill assessment tools. Methods to improve the psychometric properties of such tools must be sought if they are to be used for high-stakes evaluations.

Rater training is a process whereby raters undergo instruction on how to best evaluate trainees and produce reliable and valid scores.[12] It was developed to address the natural bias introduced by subjective performance assessments. There is compelling evidence in the behavioural and social sciences literature that rater training can improve rater performance.[12–16] In a landmark systematic review and meta-analysis on rater training, Woehr and Huffcutt[12] categorized 4 different types of training, including rater error training, behavioural observation training, performance dimension training and frame-of-reference (FOR) training. They found FOR training to be the most effective strategy to improve rater behaviour, with the other types of training having more moderate effects. Frame-of-reference training builds a common construct between raters, which they use to evaluate subjects. Studies have shown improvements in rater agreement for several observational clinical instruments following FOR training.[13–16] Rater training has not been well assessed in medical or surgical education, despite its support in other disciplines.[17–23] The reliability of the OSATS GRS and other tools may be improved with the use of rater training, but, to our knowledge, this has not been studied in any well-powered or systematic fashion. In the present study, we sought to determine the utility of rater training for surgical skills assessment. Specifically, we evaluated the effect of a brief FOR training session on the reliability of surgeon raters' evaluations of trainees performing a suturing and knot-tying task.

Methods

Study design

An advertisement requesting participants was distributed to eligible surgeons within the Winnipeg Regional Health Authority and surrounding area. Participation was voluntary, and surgeons were recruited between September 2013 and April 2014. Attending surgeons from any surgical subspecialty certified by the Royal College of Physicians and Surgeons of Canada were eligible for participation. Participating surgeons were randomly allocated to either rater training or no training by means of stratified block randomization. Surgeons were stratified by practice location and subspecialty, as these factors had a significant effect on surgeons' ratings in a previous review.[19] Trainee assessments and rater training were performed during an individual session with each surgeon and 1 of the study administrators (R.L.R.). Randomization was performed at the beginning of the session; identical unmarked envelopes with designation to either rater training or no training were used. Surgeons in the no-training group proceeded directly to evaluating trainee assessment videos, with no further instruction. Those assigned to rater training underwent FOR training in the form of a training video, which they viewed immediately before evaluating the trainee assessment videos. Participants in the no-training group could watch the training video after rating if desired. The University of Manitoba Health Research Ethics Board approved this study.

Intervention

We selected FOR training for the rater training intervention based on the findings of Woehr and Huffcutt.[12] Frame-of-reference training builds a shared understanding of rating standards among raters. This is accomplished by explicitly defining terms on the tool and giving examples of performance levels that would be expected for a given rating. We developed a 7-minute training video incorporating FOR elements, which was reviewed for face validity before the study by 3 surgeons (A.V., L.M.G. and J.P.) with graduate degrees in education. The video is available for viewing online (https://youtu.be/CzF-hEywufQ). The video included an introduction to FOR training and 3 assessment tools: a visual analogue scale, a task-specific checklist and the GRS. The visual analogue scale was described first, and the extremes of the scale were defined. The top of the scale was defined as a performance one would expect from a competent surgeon. This is theoretically the top goal of training. This was described as completion of the task smoothly, correctly and proficiently, without any major errors. The bottom end of the scale was defined as inability to complete the task independently or completion of the task with major or repeated errors in each step. Next, the task-specific checklist was described. The correct performance of the task is intrinsically included in the checklist. Incorrect performance is not given, so common errors contributing to incorrect performance for each step were defined and described. Finally, the GRS was described, with emphasis placed on the tool's being used "irrespective of training level," as stated on the form. Those evaluating trainees of all levels should rate them using the same set of criteria with this construct. The task (suturing) was broken down into several key procedural steps because the GRS applies to broad technical properties throughout the procedure. Scores were defined by the approximate number of errors in each step. Scores of 1 would have numerous errors in all steps, and scores of 5 would having errors in no steps (representing a performance that would be expected of an independent, competent surgeon). We also defined the error terms in each domain. For instance, in the "respect for tissue" domain, we defined "unnecessary force" as grasping the skin edges too roughly or jamming the needle through the tissue without following the curve.

Trainee assessment videos

We chose simple suture and instrument tie as the task for evaluation. A simple suture and knot tie facilitated recruitment of surgeons from a wide range of surgical specialties and allowed raters to evaluate multiple trainees during a single rating session. We developed 10 videos of trainees performing the task on a plastic suture model for evaluation. The trainees were selected to include a range of training levels, from third-year medical students to third-year general surgery residents. The videos showed only the trainees' gloved hands in the operative field to allow for blinded assessments. The videos were presented to raters in a random sequence. Raters watched each video once and then completed all 3 assessment tools. The process was repeated until all 10 videos had been rated.

Technical skill assessment tools

The surgeon raters used 3 separate assessment tools to evaluate trainees' technical skills: a visual analogue scale, a task-specific checklist and a modified version of the OSATS GRS[5] (Appendix 1, available at canjsurg.ca/015917-a1). The visual analogue scale consisted of a 10-cm horizontal line with verbal descriptors at either end ("Was unable to complete the task" and "Could safely and independently perform the task"). Raters marked an "X" on the line indicating the trainee's level of overall technical competence. Scores were determined by measuring the location where the mark intersected the line, for a maximum possible score of 10. We modified a 10-item task-specific checklist from a previously published checklist for a suturing and knot-tying task.[24,25] Raters gave 1 point for each correctly performed checklist item and 0 points for incorrectly or incompletely performed items. The points for each item were summed, for a maximum possible score of 10. Finally, we used a modified version of the OSATS GRS, as certain aspects of the original scale such as "knowledge of instruments" and "use of assistants" could not be evaluated on our assessment videos. The final adapted GRS consisted of 5 items: respect for tissue, time and motion, instrument handling, flow of procedure and overall performance.[3,5] Each item was rated on a 5-point Likert scale, with 5 representing the top score. We calculated an average of the scores for the 5 items, for a maximum possible score of 5.

Statistical analysis

We measured differences between the 2 rater groups using an independent samples t test for continuous variables and the χ^2 test for categorical variables. We calculated mean assessment tool scores using mixed-effects models. We measured IRR using intraclass correlation (ICC) type 2. Analyses were run with SPSS version 17.0 (SPSS Institute) and R Console version 3.1.0 (R Foundation for Statistical Computing).

Results

Participant characteristics

A total of 47 surgeons from surgical subspecialties including general surgery, urology, orthopedics, otolaryngology,

neurosurgery, thoracic surgery, cardiac surgery and plastic surgery participated, 24 in the training group and 23 in the no-training group. Characteristics for the 2 groups are listed in Table 1. There were no significant differences between the 2 groups in surgeon characteristics, practice setting, or teaching and education experience. The only exception was years in practice, with surgeons in the training group being in practice slightly longer than those in the no-training group ($p = 0.04$). There was no significant difference between the groups in the number of surgeons with previous experience with evaluation tools. None of the participants had any previous exposure to formal rater training.

Assessment tool scores

Mixed-model analysis did not show any significant effect of training on mean scores for any of the tools (Table 2). In fact, mean scores were nearly identical for the 2 groups. Rater training added only 0.02 to the mean scores for the visual analogue scale and the GRS, and resulted in a decrease in mean score of 0.41 on the task-specific checklist.

Reliability

A statistically significant improvement in IRR was not shown for any of the assessment tools (Table 3). The reliability values were higher for the training group than for the no-training group on each of the 3 assessment tools. However, the 95% confidence intervals overlapped, which made the differences nonsignificant. We performed subgroup analyses for the IRR of raters within the general surgery and academic setting cohorts, but no significant effect of training was shown within these subgroups (data not shown).

Table 1. Characteristics of surgeons who received or did not receive rater training

Characteristic	No. (%) of participants*		
	No training $n = 23$	Rater training $n = 24$	p value
Age, yr, mean ± SD	42.7 ± 7.2	46.5 ± 8.1	0.6
Years in practice, mean ± SD	9.5 ± 7.3	14.4 ± 8.2	0.04
No. of residents per year, mean ± SD	15.2 ± 15.6	14.7 ± 9.2	0.9
No. of medical students per year, mean ± SD	12.5 ± 11.4	10.9 ± 6.3	0.6
Previous experience with evaluation tools	11 (48)	10 (42)	0.8
Specialty			0.9
General surgery	17 (74)	18 (75))	
Other subspecialty	6 (26)	6 (25)	
Practice setting			1.0
Academic	19 (83)	22 (92)	
Community	4 (17)	2 (8)	0.1
University appointment	22 (96)	20 (83)	
Sex			0.8
Female	5 (22)	6 (25)	
Male	18 (78)	18 (75)	

SD = standard deviation.
*Except where noted otherwise.

Table 2. Mean assessment tool scores from mixed-model regression

Assessment tool	Mean score		
	No training	Rater training	p value
Visual analogue scale*	5.41	5.43	1.0
Task-specific checklist*	7.79	7.38	0.1
Global rating scale†	2.76	2.78	0.8

*Maximum score 10.
†Maximum score 5.

Table 3. Interrater reliability of assessment tools

Assessment tool	ICC2 (95% CI)	
	No training	Rater training
Visual analogue scale	0.46 (0.27–0.75)	0.71 (0.50–0.91)
Task-specific checklist	0.33 (0.17–0.64)	0.46 (0.27–0.75)
Global rating scale	0.61 (0.41–0.85)	0.71 (0.52–0.89)

CI = confidence interval; ICC2 = interclass correlation type 2.

Discussion

Recent changes to traditional medical education formats are presenting new challenges when training physicians. Several certification bodies have announced initiatives to move toward competency-based training models.[26–28] These developments require high-quality performance evaluations to ensure training programs are producing competent clinicians. Many programs have introduced the use of standardized clinical assessment tools to ensure these goals are met.[4] However, the reliability of many assessment tools is insufficient for high-stakes testing, and the literature often fails to examine methods to improve reliability.[20,23,29] Ways to enhance the reliability of these tools must be developed if they are to be used for purposes such as determining proficiency and advancement in competency-based frameworks. Reliability can be improved by increasing the number or variability of trainee assessments, modifying the current assessment tools or decreasing the subjectivity of raters, such as with rater training.[30] In surgical education, substantial time constraints for both trainees and evaluators make increasing the number of assessments challenging.[22] Changing existing tools creates problems pertaining to having multiple versions of a similar tool. This, at times, requires revalidation.[8] As a result, despite some limitations, rater training remains

one of the most promising means available for improving the reliability of existing validated assessment tools.

Rater training seeks to minimize bias from preexisting rater beliefs and produce more reliable scores. Several studies have shown a positive and prolonged effect of physician rater training on the reliability, accuracy and quality of medical training assessment tools.[20,23,31–33] In 2 previous studies, rater training for surgeons evaluating technical skills was assessed; neither was able to show a significant effect of training.[17,19] However, those studies were likely underpowered, and optimal training methods may not have been used.

In the current study, a significant effect of rater training on the IRR was not shown for 3 technical skill assessment tools, including the OSATS GRS.[5] For educational purposes, the ICC of 0.71 for both the GRS and visual analogue scale for the rater training group would be considered "good" agreement.[34] The ICCs of 0.61 and 0.46 for the GRS and visual analogue scale, respectively, for the untrained group would be considered only "moderate" agreement. Although there were trends toward improved reliability in the training group for all 3 tools, confidence intervals were wide and overlapping. In addition, in all cases, the IRR remained below the minimum desired threshold of 0.8 for high-stakes testing.[10,11]

There are multiple factors that may have contributed to why this investigation did not show a significant difference between the trained and untrained groups. These factors include the power of the study, the design of the rater training intervention, the assessment forms themselves, the characteristics of the raters and the complexity of the skill being assessed. Obtaining adequate power for reliability studies of this nature is considerably limited by the number of assessments performed.[35] About 60 assessments from each rater would have been needed to obtain sufficient power for ICC.[35] This would be impractical for most medical education studies, as they are often restricted by the number of trainees and evaluators participating.[22] This limitation makes adequately powered studies assessing the effect of rater training on reliability difficult to execute. One consideration is to have fewer raters perform a much larger number of assessments over a longer period. This approach may also be quite difficult to facilitate in practice, which highlights the challenge in designing robust reliability studies.

Another major challenge in studying rater training is the lack of evidence on the ideal training format, as there is nearly unlimited variability in the way training can be administered. This makes it difficult to develop a new training intervention. A variety of training formats have been described, varying from in-person tutorials to training videos, training ranging from less than an hour to multiple-day workshops, and the use of single or multiple types of rater training.[17–23,32,33] Time constraints in surgical practice are often a limiting factor to surgeon buy-in and participation. For the present study, we developed a brief FOR training video, which raters viewed before the rating session. Administering a standard video immediately before each rating session has the advantage of ensuring the same training of all raters, even if different surgeons often evaluate trainees at any given session. We felt that a longer training session was impractical and unlikely to be widely adopted given the time constrains most surgeons face in practice. It is possible that the training intervention was too brief to lead to effective changes in rater behaviour. However, it has been suggested that longer training is unlikely to be of benefit for relatively intuitive assessment tools or skills.[19] Ongoing training or multiple sessions may have been necessary to integrate the training construct.[14] The fact that 3 assessment tools were covered during the brief training session may have made it difficult for raters to remember all the information given for each tool. Focusing on only 1 tool may have led to better retention of the training information and may have decreased the effect of rater fatigue. We had hoped to determine whether the effect of training would differ between more standardized tools like the OSATS GRS and less validated tools. The GRS had higher reliability than the visual analogue scale or the task-specific checklist, which may indicate that it has an inherent quality that improves rater reliability. Despite this, the GRS still failed to achieve the desired reliability of 0.8 in either group. Training led to improvement in reliability for all 3 tools, which suggests that rater training may play a role in improving rater behaviour despite any inherent characteristics of the rating tool.

Rater characteristics may also affect the degree to which they respond to rater training. Expert raters are more likely to be resistant to rater training, especially if the constructs of training differ significantly from their own established ideas and principles.[32,36] Some raters may have fundamentally poor rating behaviours that are less likely to respond to training. Physician raters in particular have been suggested as a group that may be inherently difficult to train, with mixed results of rater training programs.[37] It has been suggested that, instead of attempting to train resistant raters, they should be identified and removed.[21] However, this is not a realistic approach when maintaining a functional academic surgical program where those willing to educate are at a premium. Although the use of untrained or nonsurgeon raters may be areas of future interest and study, these methods are unlikely to be adopted into widespread practice anytime soon. With competency-based training becoming a reality in the near future, methods to improve reliability that can be conducted by current surgeon evaluators are needed. Moreover, the only apparent downside of rater training is the potential time commitment, which, in the current study,

was quite brief. A trial of rater training is therefore still warranted, as it may prove beneficial with minimal negative effects.

We selected a simple suture and knot-tying task performed by junior trainees for evaluation. We felt that if rater training had a positive effect for a simple skill, training could subsequently be applied to more complex tasks. Owing to the innate ability of surgeons to evaluate simple tasks, with minimal variation in baseline reliability between raters, the chosen task may ultimately have been too basic and brief to show a significant effect of training. Whether basics tasks require rater training for evaluation is questionable, and the universal nature of these tasks may make it difficult to teach new evaluation paradigms during training. Increased trainee variation with a wider performance range can improve reliability.[30] Reliability may also have been increased by including a broader range of trainees or by evaluating a more specialized or complex task better suited to rater training. Finally, practising surgeons would rarely perform this many sequential assessments in such a short period. This aspect of our study design may have introduced rater fatigue and decreased reliability. Methods to study technical skill that more closely emulate real-life scenarios ultimately need to be developed. In the future, having raters assess more complex tasks over a longer time with fewer evaluations at each session may prove beneficial in improving reliability and power.

Limitations

This study was limited by statistical power and the effect size. There are several limitations inherent to the study design required of observational surgical education research. The investigational nature of the assessments may have had an effect on rater behaviour and introduced bias. Raters may have adjusted their behaviour owing to the fact they were being observed or because the evaluations were for research purposes. Finally, in practice, surgeons would rarely perform multiple sequential assessments over such a short period. This aspect of our study design may have introduced rater fatigue and decreased reliability.

Conclusion

This randomized controlled trial to assess the reliability of technical skills assessment did not show a statistically significant difference between rater groups. This result has led to reexamination of the content, length and administration of the rater training intervention. Accepting the study limits, it is interesting to note the trend toward higher ICC in the rater training group for each of the 3 assessment tools. Combined with results from other fields,[12–16] these findings support the belief that the effect of rater training is real and should be considered as a means to improve reliability of technical skills assessment. Optimal training formats, in particular for surgeon raters, have yet to be defined. To our knowledge, no studies of technical skill assessment tools involving surgeon raters have produced agreement above the desired threshold of 0.8. As surgical education moves toward competency-based training models, continued efforts to improve reliability are clearly needed, even for the most extensively studied and validated technical assessment tools.

Affiliation: From the Department of Surgery, University of Manitoba, Winnipeg, Man.

Competing interests: None declared.

Contributors: All authors designed the study. R. Robertson acquired and analyzed the data, which A. Vergis and J. Park also analyzed. R. Robertson and J. Park wrote the article, which all authors reviewed and approved for publication.

References

1. Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993; 165:358-61.
2. Hove PD Van, Tuijthof GJM, Verdaasdonk EGG, et al. Objective assessment of technical surgical skills. *Br J Surg* 2010;97:972-87.
3. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skills (OSATS) for surgical residents. *Br J Surg* 1997;84:273-8.
4. Gray JD. Global rating scales in residency education. *Acad Med* 1996; 71(1 Suppl):S55-63.
5. Reznick R, Regehr G, MacRae H, et al. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173:226-30.
6. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg* 2005;190:107-13.
7. Vassiliou MC, Kaneva PA, Poulose BK, et al. Global Assessment of Gastrointestinal Endoscopic Skills (GAGES): a valid measurement tool for technical skills in flexible endoscopy. *Surg Endosc* 2010;24:1834-41.
8. Szasz P, Louridas M, Harris KA, et al. Assessing technical competence in surgical trainees: a systematic review. *Ann Surg* 2015;261:1046-55.
9. Faulkner H, Regehr G, Martin J, et al. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med* 1996;71:1363-5.
10. Sidhu R, Grober E, Musselman L, et al. Assessing competency in surgery: Where to begin? *Surgery* 2004;135:6-20.
11. Crossingham GV, Sice PJA, Roberts MJ, et al. Development of workplace-based assessments of non-technical skills in anaesthesia. *Anaesthesia* 2012;67:158-64.
12. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occup Organ Psychol* 1994;67:189-205.
13. Cusick A, Vasquez M, Knowles L, et al. Effect of rater training on reliability of Melbourne Assessment of Unilateral Upper Limb Function scores. *Dev Med Child Neurol* 2005;47:39-45.
14. Müller MJ, Rossbach W, Dannigkeit P, et al. Evaluation of standardized rater training for the Positive and Negative Syndrome Scale (PANSS). *Schizophr Res* 1998;32:151-60.
15. Müller MJ, Dragicevic A. Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. *J Affect Disord* 2003;77:65-9.
16. Evans LV, Morse JL, Hamann CJ, et al. The development of an independent rater system to assess residents' competence in invasive procedures. *Acad Med* 2009;84:1135-43.

17. Rogers DA, Regehr G, MacDonald J. A role for error training in surgical technical skill instruction and evaluation. *Am J Surg* 2002;183: 242-5.
18. Spanager L, Beier-Holgersen R, Dieckmann P, et al. Reliable assessment of general surgeons' non-technical skills based on video-recordings of patient simulated scenarios. *Am J Surg* 2013;206:810-7.
19. George BC, Teitelbaum EN, DaRosa DA, et al. Duration of faculty training needed to ensure reliable or performance ratings. *J Surg Educ* 2013;70:703-8.
20. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med* 2004;140:874-81.
21. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. *Med Educ* 1980;14:345-9.
22. Noel GL, Herbers JE, Caplow MP, et al. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med* 1992;117:757-65.
23. Holmboe ES, Huot S, Chung J, et al. Construct validity of the mini-clinical evaluation exercise (miniCEX). *Acad Med* 2003;78:826-30.
24. Figert PL, Park AE, Witzke DB, et al. Transfer of training in acquiring laparoscopic skills. *J Am Coll Surg* 2001;193:533-7.
25. Rosser JC, Rosser LE, Savalgi RS. Skill acquisition and assessment for laparoscopic surgery. *Arch Surg* 1997;132:200-4.
26. Competence by Design. Ottawa: Royal College of Physicians and Surgeons of Canada. Available: www.royalcollege.ca/rcsite/cbd/competence-by-design-cbd-e (accessed 2016 Mar. 31).
27. Medical education explores competency-based assessment. AMA Wire 2014 July 10. Available: https://wire.ama-assn.org/ama-news/medical-education-explores-competency-based-assessment (accessed 2017 Jan. 22).
28. Competency-based training in medical education — 2010. Australian Medical Association 2010 Aug. 16. Available: https://ama.com.au/position-statement/competency-based-training-medical-education-2010 (accessed 2017 Jan. 22).
29. Kroboth FJ, Hanusa BH, Parker S, et al. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med* 1992;7:174-9.
30. Wanzel KR, Ward M, Reznick RK. Teaching the surgical craft: from selection to certification. *Curr Probl Surg* 2002;39:583-659.
31. van der Vleuten CPM, van Luyk SJ, van Ballegooijen AMJ, et al. Training and experience of examiners. *Med Educ* 1989;23:290-6.
32. Dudek NL, Marks MB, Bandiera G, et al. Quality in-training evaluation reports — Does feedback drive faculty performance? *Acad Med* 2013;88:1129-34.
33. Dudek NL, Marks MB, Wood TJ, et al. Quality evaluation reports: Can a faculty development program make a difference? *Med Teach* 2012;34:e725-31.
34. Gwet KL. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters*. 3rd ed. Gaithersburg (MD): Advanced Analytics, LLC; 2012.
35. Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Stat Med* 1998;17:1157-68.
36. Feldman M, Lazzara EH, Vanderbilt AA, et al. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof* 2012;32:279-86.
37. Ludbrook J, Marshall VR. Examiner training for clinical examinations. *Br J Med Educ* 1971;5:152-5.