Evaluating the reliability of surgical assessment methods in an orthopedic residency program

Nicholas Smith, MD, MSc
John Harnett, MB BCh
Andrew Furey, MD, MSc

Accepted for publication
Apr. 7, 2015

Correspondence to:

N. Smith
Health Sciences Centre
Department of Orthopaedics
Memorial University of Newfoundland
300 Prince Philip Dr
St John's NL A1B 3V6
nicksmith@munmed.ca

DOI: 10.1503/cjs.010614

Background: Orthopedic surgical education in Canada has seen major change in the last 15 years. Work hour restrictions and external influence have led to new approaches for surgical training. With a change toward competency-based educational models under the CanMEDS headings there is a need to ensure the validity of modern assessment methods. Our objective was to evaluate the reliability of a currently used surgical skill assessment tool within an orthopedic residency program, as measured by the Surgical Encounters Form.

Methods: A surgical assessment tool has previously been created at our institution that comprises 15 items spanning 4 of the CanMEDS competencies. Results were blinded to the primary investigator and coded by a third party. The assessments were collected, and we measured percent agreement using Cronbach's α and Fleiss κ .

Results: Over a 5-month period 11 staff members assessed 10 residents. Eighty-eight assessments were completed in total. Weighted percent agreement was 90.9%. Cronbach's α averaged 0.865 for the medical expert role, 0.920 for technical skills, 0.934 for the communicator role, 1.00 for the collaborator role and 1.00 for the health advocate role. The mean Fleiss κ score was 0.147 (95% confidence interval -0.071 to 0.364), demonstrating low interrater reliability.

Conclusion: Despite the development of a validated assessment tool to evaluate surgical skills acquisition, interrater reliability results suggest low levels of agreement among assessors.

Contexte : L'enseignement de la chirurgie orthopédique au Canada a beaucoup évolué ces 15 dernières années. La diminution des heures de travail et les influences extérieures sont à l'origine des nouvelles approches pour la formation en chirurgie. Avec l'avènement des modèles de formation fondés sur les compétences et les rôles CanMEDS, il faut s'assurer de la validité des méthodes d'évaluation modernes. Notre objectif était d'évaluer la fiabilité d'un outil de vérification des habiletés chirurgicales actuellement utilisé par un programme de résidence en orthopédie, au moyen du Formulaire de rencontres chirurgicales.

Méthodes : Un outil d'évaluation chirurgicale existe déjà dans notre établissement et comporte 15 éléments qui se rattachent à 4 des rôles CanMEDS. Sur une période de 5 mois, 11 membres du personnel ont évalué 10 résidents. Les résultats étaient à l'insu de l'investigateur principal et encodés par une tierce partie. Les évaluations ont été colligées et nous avons mesuré le pourcentage de concordance à l'aide du coefficient α de Cronbach et du score κ de Fleiss.

Résultats : Quatre-vingt-huit évaluations ont été effectuées au total. Le pourcentage de concordance pondéré a été de 90,9 %. Le coefficient α de Cronbach a été en moyenne de 0,865 pour le rôle d'expert médical, de 0,920 pour les habiletés techniques, de 0,934 pour le rôle de communicateur, de 1,00 pour le rôle de collaborateur et de 1,00 pour le rôle de promoteur de la santé. Le score κ de Fleiss moyen a été de 0,147 (intervalle de confiance de 95 % -0,071 à 0,364), témoignant d'une faible fiabilité inter-évaluateurs.

Conclusion : Malgré la mise au point d'un outil d'évaluation validé pour mesurer l'acquisition des habiletés en chirurgie, les résultats au plan de la fiabilité inter-évaluateurs indiquent un faible degré de concordance entre les examinateurs.

By 2014, the training of orthopedic surgery residents in Canada will have undergone another fundamental change. The Royal College of Physicians and Surgeons of Canada (RCPSC) is adopting a competency-based training program that all medical training institutions will have to abide by. This will build on the CanMEDS competency framework that has been in place since 1993. The framework is a “common set of essential abilities that all physicians, regardless of specialty, need for optimal patient outcomes.”¹ The 7 components of the framework are medical expert, communicator, collaborator, manager, health advocate, scholar and professional.^{2,3}

With increased public demand for accountability, government pressure, advancing technology, work hour restrictions and financial limitations there has been a shift toward defined, objective, competency-based learning.⁴ Though the term competency is often confused, it implies that aspiring residents are required to demonstrate a core set of knowledge and skills at an expected level before they can be allowed to advance and practise without restriction.⁵

The University of Toronto Department of Orthopaedics has taken a lead within the surgical community in evaluating the feasibility of a competency-based training program.⁶ They recently completed their 3-year experience with the first generation of these newly trained residents. They recognize the need for further development of the curriculum map in order to produce long-term, sustainable results. An important component of this future work will be the development of valid assessment methods.

The assessment of surgical residents is complex. A surgeon has passed through several stages of training, including medical school, residency, possibly subspecialty training and continuing professional development. At each of these phases he or she has many roles with different expectations from themselves, from the community and from their employers. Several questions arise: Who should perform surgical assessments? What are the expectations? What is the minimum standard? How does one assess technical skill? What format should be used? Which type of assessment method is best? What is the gold standard?

Each instance of assessment should be performed under the heading of one of the CanMEDS roles. Surgical education requires not only a firm knowledge base, but also a mastery of technical skills. This broadly falls under the medical expert role of CanMEDS. One of our program's assessment methods of surgical skill involves the Surgical Encounters Form (SEF; Appendix 1, available at cansurg.ca). This 15-item form incorporates 4 of the CanMEDS competencies — medical expert, communicator, collaborator and health advocate — as well as a section for technical skill in order to fully evaluate surgical competence. A consensus

panel of staff surgeons created the SEF to fulfill the CanMEDS requirements for an orthopedic residency training program. The evaluation tool has been modified several times based on feedback from staff and residents. The SEF is completed at least once during every resident rotation.

It is crucial that assessment methods of surgical residents be valid.⁷ Validity is the concept of discerning if an assessment tool is actually measuring that which it purports to measure. There is conflicting literature on the requirements for demonstrating validity, but Cook and Beckman⁸ provide a modern, medically oriented definition that is suitable and understandable.

Validity can be broken down into 5 factors: content, response process, relationship to other variables, consequences and internal structure. The content of a tool should represent the entire construct it is evaluating. There should be no extraneous information or deviation from the spirit of the construct. On the other hand, there should be no missing pertinent details. Second, the response process should demonstrate that a tool's outcomes reflect the user's thoughts during an assessment moment. Cook explains that if an evaluator or a student were to speak out loud and describe their thoughts during an assessment, the tool should adequately reflect these vital moments. If there is the possibility to be good, bad or ugly, the response process must reflect this. In essence, a valid tool must be built on foundations that reflect the mental process of the assessment. Next, any new assessment tool should be comparable to currently used methods and should most closely align with the gold standard. Similar evaluation methods should correlate with each other. The fourth factor in determining validity is the concept of consequence. Does the score make a difference? Can we take some amount of meaning from the result of the measure and take action based on the results? Ideally any type of resident evaluation tool would aid in academic advancement, job applications, guidance toward extra training and identifying areas of weakness. The final concept is the internal structure or reliability of the tool. Evaluation methods of similar items should yield similar results among users and over time. Each of the 5 components of validity is required to demonstrate the true value of a tool. Cook and Beckman⁸ recognized that the concept of validity is a fluid one. In any one instance of evaluation a tool may or may not fulfill the criteria for validity. A key component of this evaluation strategy is to build a large body of evidence in multiple situations to ensure that a reliable conclusion is being met. No single evaluation will confirm the validity of a tool.

Reliability is the concept that a measurement tool can achieve reproducible results among users and at different points in time.⁹ Terms often used synonymously with reliability are repeatability, precision and consistency. Errors in reliability can be either systematic or random and affect the validity of a study. Systematic errors occur in the same

way each time a measurement is performed. Random errors occur differently for each assessment. Reliability is an ideal surrogate measure of validity because it has defined mathematical values depending on which method of evaluation is used. If a study is not reliable, it is not valid. On the other hand if a study is reliable, it may be valid. The other components of validity are more qualitative and more difficult to demonstrate numerically.

We had previously attempted to determine the reliability of a non-medical expert, or intrinsic, CanMEDS role at our institution. Problems with feasibility, complex wording, lengthy assessment forms and poor staff compliance led to an inadequate number of responses and subsequently questionable results. Hulley and Cummings¹⁰ proposed 5 key steps to improve reliability measures of evaluations: standardize the measurement methods, train observers, refine instruments, automate instruments and take repeat measurements. Each of these recommendations was undertaken in preparation for this project.

By determining the reliability of evaluation methods created in the CanMEDS context we can move closer to determining the usefulness of this assessment scheme. The purpose of our study was to determine the reliability of a surgical assessment tool, broadly under the medical expert role, within an orthopedic surgery residency program.

METHODS

We performed a literature review in February 2013 and updated it in October 2014 using PubMed, EMBASE and Cochrane search engines and the terms [resident] + [evaluation] + [competence] + [CanMEDS] and [surgery]. We found 18 studies that address surgical assessment methods of residents (Fig. 1). The methods of assessment included objective structured assessment of technical skill (OSATS),^{11,12} structured technical skills assessment forms¹³ and individually created tools for the assessment of task-specific objectives. Each of the studies recognizes a void left in evaluation methods of technical skill and attempts to create reliable options for these assessment moments.

Our health research ethics board approved our study. In July of 2013, we held individual meetings with the staff orthopedic surgeons and the orthopedic residents. During these sessions, we explained the purpose of the study, answered questions and obtained consent. Staff surgeons were already familiar with the SEF. The 3-point grading scale was explained carefully. All comparisons were made to staff surgeons. A score of 3 reflects equivalent skill to that of a board-certified surgeon, 2 reflects capability but not quite the skill level of a staff surgeon, and 1 reflects insufficient skill. A fourth category was available for “not

Fig. 1: Literature search summary.

observed.” The surgeons completed assessments during operating days for all orthopedics residents on service. Residents off service were excluded, as were off-service residents covering the orthopedics team. Residents ranged from postgraduate year (PGY)-1 to PGY-5. The staff surgeon and resident would agree on a case for assessment during each operating day. An electronic copy of the form, which could be completed on hand-held devices, was emailed to the staff surgeon, who was encouraged to complete the form as soon as possible after the operation. The form was submitted electronically to a third party (the program research coordinator). Upon completion of the study all assessments were coded to keep the principle investigator blinded to the study results.

Statistical analysis

Data were collected and analyzed for internal consistency using Cronbach’s α and for interrater reliability using percent agreement and Fleiss κ scores. We used SPSS version 20 to evaluate internal consistency. The same data were entered into AgreeStats2013 version 2 to assess the percent agreement and Fleiss κ scores for weighted data.

We used Cicchetti’s method for ordinal data, in which the number of categories is squared and multiplied by 2, for the original sample size calculation.¹⁴ Thirty-six evaluations were required to adequately assess the interrater reliability of the tool through a weighted measurement of Fleiss κ .

RESULTS

Eleven staff members assessed 10 residents over a 5-month period. Eighty-eight evaluations were collected in total. One evaluation contained no resident identification and was discarded, leaving 87 evaluations for analysis (Table 1).

Cronbach’s α measures averaged 0.865 for the medical expert role, 0.920 for technical skills, 0.934 for the communicator role, 1.00 for the collaborator role and 1.00 for the health advocate role (Table 2).

The Agreestats2013 linear weighting scale was applied, and the average weighted percentage agreement was 0.909. The mean Fleiss κ score was 0.147 (95% confidence interval [CI] -0.071 to 0.364) for weighted data (Table 3).

DISCUSSION

The determination of reliability is a key prerequisite for the development of valid assessment tools. The RCPSC’s shift toward competency-based education will require the creation of valid assessment tools in order to uphold the fundamentals of this education strategy. Other areas of surgical practice have demanded similar scrutiny. In 2004 Furey¹⁵ sought to determine the reliability of commonly used fracture classification systems in an ortho-

pedic setting. Orthopedic literature has been flooded with classification tools for the purpose of determining prognosis and directing treatment. Without sufficient reliability these tools would lack validity, and any action taken based on the classification of fracture could be potentially harmful. Furey¹⁵ determined that for 3 commonly used fracture classifications there was low to

Table 1. Number of surgical assessments completed for each resident

Staff	Resident; no. of evaluations									
	1	2	3	4	5	6	7	8	9	10
A	—	—	—	—	—	—	3	—	—	3
B	—	—	—	—	—	—	—	4	—	2
C	2	—	—	—	6	2	—	1	—	—
D	—	—	4	—	1	3	—	1	2	—
E	—	—	1	—	5	4	—	1	—	—
F	1	—	2	—	—	2	—	2	1	1
G	2	1	1	—	3	3	—	—	1	—
H	—	1	1	—	4	2	—	—	—	—
I	—	—	—	—	1	—	—	1	—	—
J	—	—	—	1	—	—	—	—	—	—
K	—	—	2	—	—	3	—	4	—	2

Table 2. Cronbach’s α scores for the surgical encounters form

Resident	Competency; Cronbach’s α				
	Medical expert	Technical skills	Communicator	Collaborator	Health advocate
1	0.939	ID	1.00	ID	ID
3	0.757	1.00	NV	ID	ID
5	0.895	0.832	0.822	1.00	1.00
6	NV	0.818	0.915	1.00	ID
7	NV	NV	ID	ID	ID
8	0.909	0.960	1.00	1.00	1.00
9	NV	1.00	ID	ID	ID
10	0.823	0.912	ID	1.00	1.00
Mean	0.865	0.920	0.934	1.00	1.00

ID = insufficient data; NV = no variability.

Table 3. Percent agreement, Fleiss κ scores and 95% confidence intervals for surgical encounters form

Resident	Weighted scores	
	% Agreement	κ (95% CI)
1	0.674	0.188 (-0.27 to 0.645)
2	0.714	0.142 (-0.627 to 0.913)
3	0.916	0.433 (0.073 to 0.792)
5	0.818	0.022 (-0.05 to 0.095)
6	0.902	0.222 (0.082 to 0.361)
7	0.939	-0.304 (-0.614 to 0.007)
8	0.468	-0.097 (-0.195 to 0.002)
9	0.948	0.111 (-0.074 to 0.297)
10	0.841	-0.095 (-0.189 to -0.002)
Mean	0.909	0.147 (-0.071 to 0.364)

CI = confidence interval.

moderate interrater reliability. The purpose of the present project was to determine the reliability of an assessment method of the medical expert role within an orthopedic surgery residency program.

Our literature review revealed 2 modern studies that had demonstrated reliable instruments for the evaluation of surgical residents. Niitsu and colleagues¹⁶ assessed their residents' technical skills using the OSATS tool over a 3-year period. They noted a positive correlation between training year and a higher OSATS score. Gofton and colleagues¹⁷ developed the Ottawa surgical competency operating room evaluation (O-Score) at the University of Ottawa. Over a 2-phase evaluation they developed a general assessment scheme that could be applied to any procedure. They demonstrated a correlation between resident training year and improved scores and came to the conclusion that the tool was reliable and practical.

Previous attempts by our group to evaluate assessment methods of orthopedic surgery residents were met with difficulty. A lack of interest, the perception of wasted time and overly complex assessment methods were the stated reasons for poor staff compliance. Before starting the evaluation of the medical expert role, steps were taken to improve feasibility and compliance. Staff members were trained on the correct definitions and uses of the SEF. Strict definitions of each category were used. Complex wording of categories was simplified and shortened. We created an online version of the instrument that could be completed on mobile devices immediately after the observed procedure, and in doing so removed some of the potential for losing assessments and for recall bias. The new electronic form was emailed to the staff each day they operated with a resident and took approximately 3 minutes to complete. Finally our study included 2 6-week evaluation periods to increase our total number of evaluations. A return of 87 assessments represents a vast improvement in response rates from our surgeons and gives our study strength when compared with the currently available literature. Low numbers and poor response rates often hamper studies. Though no formal survey was done to determine reasons for the increased compliance, the feasibility and ease of the online tool, specifically through mobile interfaces, was noted to be a vast improvement.

Our literature review yielded several studies that explored novel evaluation methods for surgical skills acquisition but none that examined currently used methods.^{18–21} Eighty-eight evaluations were completed in our study, 87 of which were suitable for analysis. Three statistical measures of reliability were used for this analysis. The first is Cronbach's α . This measure seeks to determine if similar items within a matrix produce similar outcomes. The SEF uses 4 of the CanMEDS competencies to evaluate surgical competence. To ensure that heterogeneity did not falsely affect the

results, each of the separate competencies' α scores was determined (Table 2). Significant numbers of "not observed" values and data that had no variability made some scores unattainable. The average α score for the medical expert role was 0.865 and for the technical skills section was 0.920. The communicator, collaborator and health advocate roles had values of 0.934, 1.00 and 1.00, respectively. An α of 0.865 represents almost perfect agreement (Table 4). Caution must be taken in analyzing the final 4 α values. Such high scores likely represent a lack of variability within the tool and, though concordant, may not be reliable.²²

The percent agreement is simply the number of times different raters agreed on the measurement. The weighted percent agreement for this study was 91%, which supports a high interrater agreement but did not take chance into consideration. With a 3-point scale there is a 33% probability that evaluators will agree on chance alone. In order to assess this we used weighted Fleiss κ scores. The mean κ score was 0.147 (95% CI –0.071 to 0.364). This demonstrates slight agreement among users when chance is considered (Table 5).

Any measurement tool that demonstrates poor reliability will have questionable validity. The reliability of the SEF is questionable. High values of percent agreement and Cronbach's α would seem to support its reliability, but low weighted κ scores suggest a less robust measure of reliability.

The response rate during this study was greatly improved from our original evaluation of assessment measures. Staff members commented that the electronic form and a less wordy assessment scheme made the tool more feasible to complete. This led to a sample size more than double the size required.

Table 4. Commonly accepted values of Cronbach's α

Cronbach's α	Internal consistency
> 0.9	Excellent
0.8–0.9	Good
0.7–0.8	Acceptable
0.6–0.7	Questionable
0.5–0.6	Poor
< 0.5	Unacceptable

Table 5. Commonly accepted values of κ

κ	Strength of agreement
< 0	Poor; less than chance
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.0	Almost perfect

Limitations

Despite the large sample size, there were still weaknesses in our study. First, although there was a high percent agreement among raters, the κ scores were quite low. Feinstein and Cicchetti²³ raised concerns about Cronbach's α and κ statistics as measures of reliability. This phenomenon is more likely to occur with narrow Likert scales. Only 3 response options were available on the SEF. Though this simplified the tool and improved feasibility, it may have harmed the objective measures of reliability and may not accurately reflect all the potential outcomes for a resident evaluation. A high agreement based on chance alone would be expected. Other statistical measures may be needed to address these issues. Future work will have to account for balance between feasibility and overly complex but more reliable tools. Second, no formal follow-up sessions were held with the staff and residents to address their qualitative concerns with the SEF. This could be added in the future as we seek to build evidence for the validity of our institution's assessment tools. Hulley and Cummings¹⁰ methods for improvement of reliability scores should be applied again for future studies.

CONCLUSION

The SEF demonstrates questionable reliability for assessing the medical expert role in an orthopedic residency program. Further modifications need to be made before it can be reliably applied to a competency-based education system. This project is valuable in that it builds on the evidence that is needed to support the validity of our assessment methods. We have also explored successful options for overcoming some of the pitfalls that may be encountered during the evaluation of assessment methods within surgical training programs.

Affiliations: From the Department of Surgery (Smith), the Faculty of Medicine, Health Sciences Centre, (Harnett) and the Department of Orthopaedic Traumatology (Furey), Memorial University of Newfoundland, St. John's, NL.

Competing interests: None declared.

Contributors: N. Smith and A. Furey designed the study. N. Smith acquired the data, which all authors analyzed. N. Smith wrote the article, which all authors reviewed and approved for publication.

References

1. Frank JR, Langer B. Collaboration, communication, management, and advocacy: teaching surgeons new skills through the CanMEDS project. *World J Surg* 2003;27:972-8.
2. CanMEDS 2005 Framework. Royal College of Physicians and Surgeons of Canada. Available: www.royalcollege.ca/portal/page/portal/rc/canmeds/framework (accessed 2012 Mar. 4).
3. Ortwein H, Knigge M, Rehberg B, et al. Validation of core competencies during residency training in anaesthesiology. *Ger Med Sci* 2011; 9:Doc23.
4. Grantcharov TP, Reznick RK. Training tomorrow's surgeons: What are we looking for and how can we achieve it? *ANZ J Surg* 2009; 79:104-7.
5. Parent F, Jouquan J, de Kettle JM. CanMEDS and other "competency and outcome-based approaches" in medical education: clarifying the ongoing ambiguity. *Adv Health Sci Educ Theory Pract* 2013; 18:115-22.
6. Ferguson PC, Kraemer W, Nousiainen M, et al. Three-year experience with an innovative, modular competency-based curriculum for orthopaedic training. *J Bone Joint Surg Am* 2013;95:e166.
7. Chou S, Cole G, McLaughlin K, et al. CanMEDS evaluation in Canadian postgraduate training programmes: tools used and programme director satisfaction. *Med Educ* 2008;42:879-86.
8. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:116.e7-16.
9. Higgins PA, Straub AJ. Understanding the error of our ways: mapping the concepts of validity and reliability. *Nurs Outlook* 2006;54:23-9.
10. Hulley SB, Cummings SR. *Designing clinical research: an epidemiological approach*. Williams and Wilkins, Baltimore. 1988: 31-42.
11. Goff B, Mandel L, Lentz G, et al. Assessment of resident surgical skills: Is testing feasible? *Am J Obstet Gynecol* 2005;192:1331-8.
12. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273-8.
13. Winkel CP, Reznick RK, Cohen R. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg* 1994;167:423-7.
14. Cicchetti DV. Testing the normal approximation and minimal sample size requirements of weighted kappa when number of categories is large. *Appl Psychol Meas* 1977;5:101-4.
15. Furey AJ. The utility of classification systems in orthopaedic surgery. St. John's (NL): Memorial University of Newfoundland; 2004.
16. Niitsu H, Hirabayashi N, Yoshimitsu M, et al. Using the objective structured assessment of technical skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today* 2013;43:271-5.
17. Gofton WT, Dudek NL, Wood TJ, et al. The Ottawa surgical competency operating room evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med* 2012;87:1401-7.
18. Reznick R, Regehr G, MacRae H. Testing technical skills via and innovative "bench station" examination. *Am J Surg* 1997;173:226-30.
19. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273-8.
20. Laeeq K, Bhatti NI, Carey JP, et al. Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope* 2009;119: 2402-10.
21. Hopmans CJ, Hoed PT, Lijckle L, et al. Assessment of surgery residents' operative skills in the operating theatre using a modified objective structured assessment of technical skills (OSATS): a prospective multi-center study. *Surgery* 2014;156:1078-88.
22. Schmitt N. Uses and abuses of coefficient alpha. *Psychol Assess* 1996;8:350-3.
23. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. *J Clin Epidemiol* 1990;43:543-9.