Passing a Technical Skills Examination in the First Year of Surgical Residency Can Predict Future Performance

Sandra de Montbrun, MD, FRCSC
Marisa Louridas, MD
Teodor Grantcharov, MD, FACS

ABSTRACT

Background The ability of an assessment to predict performance would be of major benefit to residency programs, allowing for early identification of residents at risk.

Objective We sought to establish whether passing the Objective Structured Assessment of Technical Skills (OSATS) examination in postgraduate year 1 (PGY-1) predicts future performance.

Methods Between 2002 and 2012, 133 PGY-1 surgery residents at the University of Toronto (Toronto, Ontario, Canada) completed an 8-station, simulated OSATS examination as a component of training. With recently set passing scores, residents were assigned a pass/fail status using 3 standards setting methods (contrasting groups, borderline group, and borderline regression). Future in-training performance was compared between residents who had passed and those who failed the OSATS, using in-training evaluation reports from resident files. A Mann-Whitney *U* test compared performance among groups at PGY-2 and PGY-4 levels.

Results Residents who passed the OSATS examination outperformed those who failed, when compared during PGY-2 across all 3 standard setting methodologies ($P < .05$). During PGY-4, only the contrasting groups method showed a significant difference ($P < .05$).

Conclusions We found that PGY-1 surgical resident pass/fail status on a technical skills examination was associated with future performance on in-training evaluation reports in later years. This provides validity evidence for the current PGY-1 pass/fail score, and suggests that this technical skills examination may be used to predict performance and to identify residents who require remediation.

Introduction

Competency-based surgical education is gaining momentum around the world due to its aim to ensure that surgeons achieve the necessary skills to provide safe patient care.[1–4] The ability to predict competence would have major implications on resident selection, promotion, and certification.[5]

While a surgeon is expected to achieve competence in several domains, technical skills remain a key component for surgical specialties. Simulated environments have been used for technical skills training and have demonstrated transferability of skills to the operating room.[6,7] However, simulated performance data to date have not been used to predict performance. Among the tools to assess technical skills,[8] 1 of the most widely used is the Objective Structured Assessment of Technical Skills (OSATS),[9] which has been implemented across a variety of specialties.[9–12] One of the limitations of the original OSATS examination was its lack of a pass score, limiting its use in pass/fail decisions.[13] Furthermore, there were no data, to our knowledge, investigating the predictive ability of this examination. Recently, pass scores have been set for the original OSATS examination, allowing residents to be assigned a pass/fail status.[14] That status used data from 513 postgraduate year 1 (PGY-1) surgical residents collected over a 10-year period to set the pass score for the OSATS examination with 3 standard setting methodologies (contrasting groups [CG], borderline group [BG], and borderline regression [BR]).[14]

One way to build further validity evidence for the OSATS examination is to demonstrate the predictive ability of the recently set OSATS pass score.[9,15] If passing or failing the OSATS examination predicts future residency performance, it not only builds validity evidence for the pass scores but also, from a practical standpoint, it could help in the early identification and remediation of underperforming trainees.

To that end, the purpose of this study was to build evidence of validity for the recently set OSATS passing scores, hypothesizing that passing the OSATS examination predicts improved future technical skills of surgical residents.

DOI: http://dx.doi.org/10.4300/JGME-D-16-00517.1

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-27 via free access

Methods

The University of Toronto (Ontario, Canada) has administered the OSATS examination to all PGY-1 surgical residents since the early 2000s. Data have been collected from all surgical residents who have taken this 8-station, simulation-based, technical skills examination since its initiation. Only raw scores have been assigned, as a passing score had not been set until recently.

A recent study used this database to set passing scores for the OSATS examination with 3 standard setting methodologies: the CG method, the BG method, and the BR method.[14] Passing scores were then used to retrospectively assign a pass/fail status to all general surgery residents (N = 133) who had taken the OSATS examination between 2002 and 2012.[14]

The current study used the pass/fail status of the 133 surgery residents to compare future residency performance between those who had passed and those who failed the OSATS. Future performance was assessed using retrospectively collected, in-training evaluation reports (ITERs) from residents' training files, capturing data from their PGY-2 and PGY-4. The ITER data were collected from all surgical rotations and completed by multiple raters.

While the ITERs include data on multiple domains of competence, our study used only data specific to technical skills with items rated on a 5-point Likert scale. A technical skills score was established for each resident during his or her PGY-2 and PGY-4 by calculating a mean score out of 5 from all of the technical skills points on their PGY-2 and PGY-4 ITERs.

A Mann-Whitney *U* test compared the technical skills score during PGY-2 and PGY-4 between residents who passed and residents who failed the OSATS using the 3 standard setting methodologies.

The Research Ethics Board at St Michael's Hospital (Toronto, Ontario, Canada) approved this study.

What was known and gap
The ability to predict future performance would allow for the early identification of residents at risk, yet most current assessments do not offer this capability.

What is new
A study assessed whether an Objective Structured Assessment of Technical Skills in the first year could predict future performance.

Limitations
Single specialty, retrospective study may limit generalizability.

Bottom line
Pass/fail status on the technical skills examination was associated with later performance on the in-training evaluation and could be used to identify residents who would benefit from early interventions.

Results

Data from the ITERs were available on 109 PGY-2s and 76 PGY-4s. The Kolmogorov-Smirnov test of normality demonstrated a deviation from normal ($P < .05$); therefore, the nonparametric Mann-Whitney *U* test was used.

The majority of PGY-2s (n = 63, 58%) had data from 2 ITERs (range, 1–3) and the majority of PGY-4s (n = 63, 83%) had 2 or 3 ITERs (range, 1–4). Each ITER contributed multiple data points for calculating a PGY-2 and PGY-4 technical skills score, respectively.

At PGY-2, a statistically significant difference was seen between residents who passed and those who failed the OSATS, according to all 3 standard-setting methods (CG, BG, BR). Those who passed outperformed those who failed (Mann-Whitney *U* test; CG, $z = 3.49$, $P < .001$; BG, $z = 2.50$, $P = .012$; BR, $z = 2.09$, $P = .037$; TABLE 1). At the PGY-4 level, this statistically significant difference was still present using the CG method (Mann-Whitney *U* test; $z = 2.58$, $P = .010$; TABLE 2; FIGURE).

Discussion

This study demonstrates that PGY-1 residents' pass/fail status on the OSATS has the potential to predict future performance, with failing residents being more likely to underperform based on ITER data during their PGY-2. As time passes, the ability to predict performance becomes more difficult, as more variables influence outcomes; despite that, the pass/fail

TABLE 1
Comparison of Postgraduate Year 2 (PGY-2) Technical Skills Scores Between Residents Who Passed and Failed the OSATS

Standard Setting Methodology	PGY-2 Overall Technical Skill Score				
	OSATS Fail		OSATS Pass		
	Median (IQR)	No.	Median (IQR)	No.	*P* Value
Contrasting groups	3.83 (0.75)	23	4.22 (0.50)	86	< .001
Borderline group	3.89 (0.76)	17	4.19 (0.49)	92	.012
Borderline regression	3.94 (0.84)	18	4.17 (0.44)	91	.037

Abbreviations: OSATS, Objective Structured Assessment of Technical Skills; IQR, interquartile range.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-27 via free access

TABLE 2
Comparison of Postgraduate Year 4 (PGY-4) Technical Skills Scores Between Residents Who Passed and Failed the OSATS

Standard Setting Methodology	PGY-4 Overall Technical Skill Score				
	OSATS Fail		OSATS Pass		
	Median (IQR)	No.	Median (IQR)	No.	*P* Value
Contrasting groups	3.67 (1.29)	20	4.10 (0.71)	56	.010
Borderline group	3.72 (1.39)	14	4.04 (0.73)	62	.17
Borderline regression	3.78 (1.33)	15	4.00 (0.74)	61	.24

Abbreviations: OSATS, Objective Structured Assessment of Technical Skills; IQR, interquartile range.

status using the CG methodology continued to predict performance in PGY-4, showing a statistically significant difference between groups. The loss of statistical significance in PGY-4 for the BG and BR methods does not discount them as useful or credible standard-setting methods; rather, this study highlights the limitation of distant prediction and the need for continuous assessment throughout training.

Progression within a surgery program often relies on ITER evaluations, which are poor at identifying residents with below-average technical skills.[16] Implementing an objective assessment of technical skills early in surgical training may be instrumental in identifying underperformers and introducing early educational interventions for effective remediation, and could help to address the failure to fail phenomenon.[17,18]

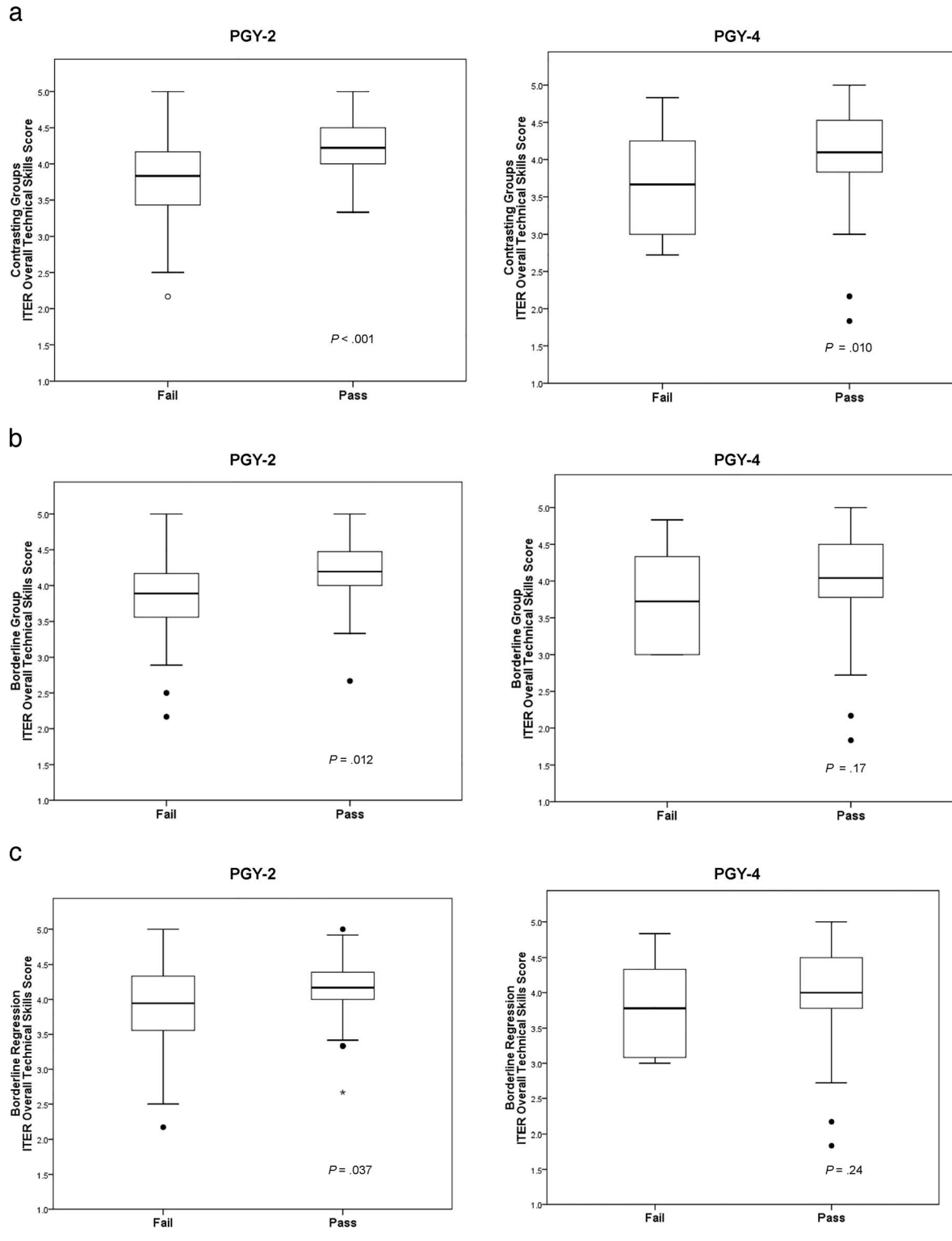
The OSATS examination, originally developed as a technical skills assessment,[19] was used in the present study to investigate the predictive ability of an objective, standardized, performance-based assessment. However, while the focus was on surgery trainees, the results of this study could be of interest to a broader surgical audience, as the OSATS has been widely adopted across other surgical specialties[8,13,20,21] and anesthesiology.[22] Furthermore, the OSATS, as a performance-based assessment, parallels the objective structured clinical examination,[23] which assesses clinical skills and has been used in nontechnical specialties, including internal medicine[24,25] and family medicine.[26] While this study focused on surgery, it provides foundational work for further predictive studies in other technical and nontechnical specialties.

Previous reports suggested ITERs are poor at identifying residents with below-average technical skills.[16] In contrast, our results suggest that ITER scores can discriminate resident technical performance. We found that a failing score on the OSATS in PGY-1 was associated with significantly poorer technical skills in PGY-2, based on ITER data. This difference was maintained in PGY-4 using the CG methodology. The absolute difference in median ITER scores, however, was small. The median ITER scores for failing residents ranged from 3.83 to 3.94 in PGY-2 and from 3.67 to 3.78 in PGY-4. In contrast, the median ITER scores for passing residents ranged from 4.17 to 4.22 in PGY-2 and from 4.00 to 4.10 in PGY-4. This suggests that a score of 3 (scale midpoint), with a descriptor of competent, may be overestimating performance at that level. This rightward shift of the assessment scale is consistent with the existing literature that ITER data are typically heavily biased toward competent. Despite that bias, the present study was still able to show a difference in ITER scores between groups. Given that ITER evaluations are already well established in many training programs, it is important to recognize this upward shift when interpreting an individual resident's ITER.

In contrast to the ITER, the OSATS has accrued a wealth of validity evidence for the interpretation of its scores.[8,13,19] However, its use in high-stakes decisions has been limited due to the lack of an established passing score.[13] Setting pass scores and investigating the effect of pass/fail status addresses the "implication or decisions" component of the Kane[27] model of validity. This domain of validity has been neglected in the OSATS validation literature and is an essential component if the OSATS is to be considered for high-stakes decisions.[9] Until recently, few studies have addressed the issue of pass/fail scores for OSATS type of examinations, typically, with a pass/fail decision based on overall dichotomous pass/fail judgment, rather than by applying standard-setting methodologies.[12,28,29] Moreover, no study, to our knowledge, has looked at the implications of OSATS pass/fail results.[9] The present study builds on the implication or decisions validity argument by demonstrating the predictive ability of the OSATS pass/fail status on future performance; this not only builds validity evidence for the OSATS but also builds validity evidence for the recently set pass scores. This component of validity is also essential for considering the use of OSATS in high-stakes assessments, such as promotion or certification.[20]

The use of technical skills simulation to assess and predict future performance in the workplace is a relatively new concept. Traditionally, simulation has

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-27 via free access



FIGURE

Comparing Technical Skills Scores at Postgraduate Year 2 (PGY-2) and PGY-4 Levels Between Residents Who Passed and Failed the OSATS Examination During PGY-1

Note: Determined with (a) a contrasting groups method; (b) a borderline group method; and (c) a borderline regression method.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-27 via free access

been used as an adjunct to *teach* technical skills, flattening the learning curve inside the operating room with studies demonstrating the transfer of skills acquired in the laboratory to the operating room.[6,7] However, data on simulation to *assess* and predict performance are limited. Moore et al,[30] used a simulated technical skills assessment during residency selection to predict performance during residency, demonstrating a moderate correlation, but did not use a dichotomous pass/fail status, limiting the ability to identify a failing cohort that would be at risk of future difficulties. The advantage of the present study is its ability to dichotomize the group into passing and failing cohorts using evidence-based passing scores, allowing for the identification of the group that would benefit from early remediation.

This study has 2 limitations. One is the use of ITER data, which have been criticized for being poor at identifying below-average residents.[16] However, while the reliability of the ITER can be low with a single rater and a single evaluation, aggregated ITER data (as used in our study) with multiple evaluators and across multiple rotations have been shown to have good reliability and predictive validity.[31–33] The second limitation is its retrospective nature. Future research will explore the ability of a pass/fail status to predict intraoperative performance and patient outcomes. Further work will also include the development of remedial strategies for underperforming residents.

Conclusion

This study demonstrated the ability of a simulated performance-based assessment to predict future skills. A key implication of these findings is the potential for early identification and remediation of the underperforming resident.

References

1. Sonnadara RR, Mui C, McQueen S, et al. Reflections on competency-based education and training for surgical residents. *J Surg Ed*. 2014;71(1):151–158.
2. Alman BA, Ferguson P, Kraemer W, et al. Competency-based education: a new model for teaching orthopaedics. *Instr Course Lect*. 2013;62:565–569.
3. Royal College of Physicians and Surgeons of Canada. Competence by design: the rationale for change. http://www.royalcollege.ca/rcsite/cbd/rationale-why-cbd-e. Accessed April 24, 2017.
4. Accreditation Council for Graduate Medical Education. Milestones. http://www.acgme.org/acgmeweb/tabid/430/ProgramandInstitutionalAccreditation/NextAccreditationSystem/Milestones.aspx. Accessed March 30, 2017.
5. Southgate L, Hays RB, Norcini J, et al. Setting performance standards for medical practice: a theoretical framework. *Med Educ*. 2001;35(5):474–481.
6. Vanderbilt AA, Grover AC, Pastis NJ, et al. Randomized controlled trials: a systematic review of laparoscopic surgery and simulation-based training. *Glob J Health Sci*. 2015;7(2):310–327.
7. Zendejas B, Brydges R, Hamstra SJ, et al. State of the evidence on simulation-based training for laparoscopic surgery: a systematic review. *Ann Surg*. 2013;257(4):586–593.
8. van Hove PD, Tuijthof GJ, Verdaasdonk EG, et al. Objective assessment of technical surgical skills. *Br J Surg*. 2010;97(7):972–987.
9. Hatala R, Cook DA, Brydges R, et al. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Health Sci Educ Theory Pract*. 2015;20(5):1149–1175.
10. Argun OB, Chrouser K, Chauhan S, et al. Multi-institutional validation of an OSATS for the assessment of cystoscopic and ureteroscopic skills. *J Urol*. 2015;194(4):1098–1105.
11. de Montbrun SL, Roberts PL, Lowry AC, et al. A novel approach to assessing technical competence of colorectal surgery residents: the development and evaluation of the Colorectal Objective Structured Assessment of Technical Skill (COSATS). *Ann Surg*. 2013;258(6):1001–1006.
12. Goff BA, Nielsen PE, Lentz GM, et al. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. *Am J Obstet Gynecol*. 2002;186(4):613–617.
13. Reznick RK, MacRae H. Teaching surgical skills—changes in the wind. *N Engl J Med*. 2006;355(25):2664–2669.
14. de Montbrun S, Satterthwaite L, Grantcharov TP. Setting pass scores for assessment of technical performance by surgical trainees. *Br J Surg*. 2016;103(3):300–306.
15. Cizek GJ, Bunch MB. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: SAGE Publications Ltd; 2007.
16. Feldman LS, Hagarty SE, Ghitulescu G, et al. Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *J Am Coll Surg*. 2004;198(1):105–110.
17. Chole RA, Ogden MA. Predictors of future success in otolaryngology residency applicants. *Arch Otolaryngol Head Neck Surg*. 2012;138(8):707–712.
18. Nadeem M, Effendi MS, Hammad Ather M. Attrition in surgical residency programmes: causes and effects. *Arab J Urol*. 2014;12(1):25–29.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-27 via free access

19. 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.
20. de Montbrun S, Roberts PL, Satterthwaite L, et al. Implementing and evaluating a national certification technical skills examination: the colorectal objective structured assessment of technical skill. *Ann Surg*. 2016;264(1):1–6.
21. Faulkner H, Regehr G, Martin J, et al. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med*. 1996;71(12):1363–1365.
22. Berkenstadt H, Ziv A, Gafni N, et al. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. *Anesth Analg*. 2006;102(3):853–858.
23. Reznick R, Smee S, Rothman A, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. *Acad Med*. 1992;67(8):487–494.
24. Daniels VJ, Bordage G, Gierl MJ, et al. Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an internal medicine residency OSCE. *Adv Health Sci Educ Theory Pract*. 2014;19(4):497–506.
25. Cruzeiro M, Bollela V. Faculty development of an OSCE in an internal medicine clerkship. *Med Educ*. 2014;48(5):545–546.
26. Skinner BD, Newton WP, Curtis P. The educational value of an OSCE in a family practice residency. *Acad Med*. 1997;72(8):722–724.
27. Kane M. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1–73.
28. Goff BA, Lentz GM, Lee D, et al. Development of a bench station objective structured assessment of technical skills. *Obstet Gynecol*. 2001;98(3):412–416.
29. Goff B, Mandel L, Lentz G, et al. Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol*. 2005;192(4):1331–1338; discussion 1338–1340.
30. Moore EJ, Price DL, Van Abel KM, et al. Still under the microscope: can a surgical aptitude test predict otolaryngology resident performance? *Laryngoscope*. 2015;125(2):E57–E61.
31. Carline JD, Paauw DS, Thiede KW, et al. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. *J Gen Intern Med*. 1992;7(5):506–510.
32. Littlefield J, Paukert J, Schoolfield J. Quality assurance data for residents' global performance ratings. *Acad Med*. 2001;76(suppl 10):102–104.
33. Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? a study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med*. 2013;88(10):1539–1544.

All authors are with the Department of Surgery, University of Toronto, Toronto, Ontario, Canada. **Sandra de Montbrun, MD, FRCSC,** is Assistant Professor of Surgery, and Colorectal Surgeon, Division of General Surgery, St. Michael's Hospital, Toronto, Ontario, Canada; **Marisa Louridas, MD,** is Senior General Surgery Resident; and **Teodor Grantcharov, MD, FACS,** is Professor of Surgery, and Minimally Invasive Surgery and Bariatric Surgeon, Division of General Surgery, St. Michael's Hospital.

Funding: This research was partly funded by the Society for Surgery of the Alimentary Tract Career Development Award for Clinical/Outcomes/Education Research.

Conflict of interest: The authors declare they have no competing interests.

These results were presented at the American College of Surgeons Clinical Congress, Chicago, Illinois, October 2015. These data were previously published as part of that meeting, as an abstract in the *Journal of the American College of Surgeons*: de Montbrun S, Grantcharov T. Passing the Objective Structured Assessment of Technical Skills (OSATS) examination predicts future technical skills performance in surgical trainees. *J Am Coll Surg*. 2015;221(suppl 4):53–54.

Corresponding author: Sandra de Montbrun, MD, FRCSC, St Michael's Hospital, University of Toronto, Room 16-064, 30 Bond Street, Toronto, ON M5B 1W8 Canada, 416.864.6060, fax 416.864.3049, demontbrunsa@smh.ca

Received August 20, 2016; revision received February 2, 2017; accepted February 21, 2017.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-27 via free access