Development and evaluation of the General Surgery Objective Structured Assessment of Technical Skill (GOSATS)

Y. Halwani[1], A. K. Sachdeva[4], L. Satterthwaite[3] and S. de Montbrun[1,2]

[1]Department of Surgery, University of Toronto, [2]Division of General Surgery, St Michael's Hospital, and [3]University of Toronto, Surgical Skills Centre, Mount Sinai Hospital, Toronto, Ontario, Canada, and [4]American College of Surgeons, Chicago, Illinois, USA

Correspondence to: Dr S. de Montbrun, Division of General Surgery, St Michael's Hospital, University of Toronto, Room 16-064, 30 Bond Street, Toronto, Ontario, Canada, M5B 1W8 (e-mail: demontbrunsa@smh.ca)

Background: **Technical skill acquisition is important in surgery specialty training. Despite an emphasis on competency-based training, few tools are currently available for direct technical skills assessment at the completion of training. The aim of this study was to develop and validate a simulated technical skill examination for graduating (postgraduate year (PGY)5) general surgery trainees.**

Methods: **A simulated eight-station, procedure-based general surgery technical skills examination was developed. Board-certified general surgeons blinded to the level of training rated performance of PGY3 and PGY5 trainees by means of validated scoring. Cronbach's α was used to calculate reliability indices, and a conjunctive model to set a pass score with borderline regression methodology. Subkoviak methodology was employed to assess the reliability of the pass–fail decision. The relationship between passing the examination and PGY level was evaluated using χ^2 analysis.**

Results: **Ten PGY3 and nine PGY5 trainees were included. Interstation reliability was 0·66, and inter-rater reliability for three stations was 0·92, 0·97 and 0·76. A pass score of 176·8 of 280 (63·1 per cent) was set. The pass rate for PGY5 trainees was 78 per cent (7 of 9), compared with 30 per cent (3 of 10) for PGY3 trainees. Reliability of the pass–fail decision had an agreement coefficient of 0·88. Graduating trainees were significantly more likely to pass the examination than PGY3 trainees ($\chi^2 = 4{\cdot}34$, $P = 0{\cdot}037$).**

Conclusion: **A summative general surgery technical skills examination was developed with reliability indices within the range needed for high-stakes assessments. Further evaluation is required before the examination can be used in decisions regarding certification.**

Presented to a clinical congress of the American College of Surgeons, San Diego, California, USA, October 2017; published in abstract form as *J Am Coll Surg* 2017; **225**: S173–S174.

Paper accepted 13 August 2019

Published online 6 October 2019 in Wiley Online Library (www.bjs.co.uk). **DOI:** 10.1002/bjs.11359

Introduction

Competency-based medical education has been embedded in health professional training internationally. The Accreditation Council for Graduate Medical Education (ACGME) and the Royal College of Physicians and Surgeons of Canada have endorsed a competency-based approach to training and assessment[1–3]. Effective from 1 July 2017, selected Canadian specialty training programmes adopted a new outcomes-based approach in the design, implementation, assessment and evaluation of programmes using the CanMEDS 2015 competency framework. This change will be adopted by general surgery training programmes by 2019, and in all disciplines by 2021[4]. In 2012, the ACGME[1] introduced the Next Accreditation System, with a shift toward competency-based medical education and development of outcome-based milestones based on the six ACGME core competencies. Similar frameworks have been implemented in Australia, and parts of Europe, including the UK[5–7]. The Union Européenne des Médecins Spécialistes[8] has worked towards harmonization of medical competence at the European level by introducing competence-based European curricula for each Specialist Section. This shift toward competency-based education will require all specialty training programmes to teach and assess trainees across the defined competency domains during their surgical training and before graduation.

Direct summative assessment of technical skill at the completion of surgical training is currently not a requirement. Commonly used methods to assess a trainee's technical skills, such as direct observation, log books and in-training evaluation reports, have problems relating to validity and reliability[5–7]. They have been criticized for lack of accuracy and objectivity, and different rater biases[9–11]. With the new era of competency-based training, specialty training programmes and professional organizations are being challenged to develop technical skills assessment methods that produce results with substantial validity evidence.

The aim of this study was to develop a simulated technical skills examination for graduating general surgery trainees, and to evaluate trainees' performance with data capture guided by Messick's validity framework[12].

Methods

In North America, following completion of medical school, applicants may enter directly into specialty training programmes in general surgery which typically last 5 years. Currently, to be eligible for certification and independent practice, Canadian surgical trainees are required to complete a 5-year Royal College-accredited surgical training programme (postgraduate years (PGYs) 1–5) after their undergraduate medical training and must successfully complete their subspecialty Royal College oral and written board examinations. Along with the final in-training reports, these examinations evaluate the trainee's medical knowledge and many of the non-technical skills critical to being a competent surgeon, such as communication and surgical decision-making.

Ethics approval was obtained from the St Michael's Hospital research ethics board. The study comprised two phases. The first phase involved examination development, and the second consisted of administration of the examination, data collection and analysis.

Phase 1: examination development

Examination creation included: development of simulated bench top models with accompanying station stem and task-specific checklist; and development of a global rating scale (GRS).

Simulated models

Eight simulated models were developed. Technical tasks were selected from a recently published Delphi consensus, outlining a blueprint for a certifying general surgery technical skills examination[13]. Operative tasks were selected from various general surgery content domains outlined in the blueprint representing the broad range of technical skills required of a practising general surgeon. Content domains included: upper gastrointestinal, lower gastrointestinal, hepatopancreatobiliary, trauma and emergency, breast, hernia, perianal and soft tissue operations. Simulated models were then developed in the surgical skills laboratory using synthetic and porcine material. Once completed, the models were trialled by two board-certified general surgeons and three senior general surgery trainees not affiliated to the study to ensure content validity and feasibility within the time constraints. Feedback was used to modify the stations accordingly.

Task-specific checklist

A task-specific checklist was developed for each model, and consisted of observable actions to be marked as 'done correctly' or 'not done/incorrect'. Each checklist was reviewed by board-certified general surgeons and senior general surgery trainees. Checklist items were reviewed to ensure that each item could be assessed within the context of the simulated model.

Global rating scale

A GRS with significant validity evidence[14] was modified to ensure that the scale was specific to the practice of general surgery, and used to capture data on six dimensions of operative performance. Each dimension of operative performance was marked on a five-point Likert scale (from 1 to 5); each point was anchored by specific descriptors, with a score of 3 describing a candidate who 'can adequately perform this procedure in independent practice'.

The GRS was used for the analysis, as it has been shown to have superior reliability to the task-specific checklist[15]. Furthermore, GRS data are better suited for high-stakes examinations[16]. A candidate's overall examination score was the sum of the GRS scores for the eight stations.

Overall skill scale

Candidates' overall performance on each station was also evaluated using an overall skill score, ranging from 1 to 5 on a Likert scale, with specific descriptors for each point. The overall skill score provided a global gestalt of each candidate's performance at that station, and was necessary for the borderline regression standard setting.

Phase 2: examination administration, data collection and analysis

Examination structure

The examination comprised ten, 15-min stations (8 technical skills stations and 2 rest stations). Eight stations were selected based on literature suggesting that eight observations provide a reliable indicator of performance[17]. At each station, a stem was provided, which outlined the clinical scenario and instructions for the task. Candidates were allowed 3 min to review the station stem and 12 min to complete the technical task. The total examination duration was 2·5 h. There were two administrations of the examination over 1 day, which took place at the University of Toronto's surgical skills centre.

Participants

Third-year (PGY3) and final-year (PGY5) postgraduate training general surgery trainees from accredited Canadian General Surgery training programmes across the country were invited to participate.

Examiners

Board-certified general surgeons from accredited surgery training programmes across the country were recruited to serve as examiners. Many of the surgeons were current or past general surgery programme directors. A total of 11 examiners were recruited, one for each technical skill station and three extra examiners whose data were used to calculate the inter-rater reliability of three stations. Examiners remained at the same station for the duration of the examination.

Validity evidence

Descriptive statistics and box plots were used to examine the performance of the PGY3 and PGY5 groups. Messick's validity framework[12] was then used to guide the accrual of validity evidence, in order to answer the specific research questions.

Internal structure evidence

Interstation and inter-rater reliability of the examination was calculated using Cronbach's α. Interstation reliability is a measure of internal consistency and reflects how consistent the participants' scores are across stations.

Consequences evidence

A pass score for the examination was set using borderline regression methodology[18]. With this method, a linear regression model is used to plot the GRS score (dependent variable) against the overall skill score (independent variable). An overall skill score of 3, which represents the borderline candidate, was inserted into the linear equation to extrapolate the corresponding GRS score, which determined the predicted GRS pass score for each station. The pass score for the examination as a whole was the sum of the pass scores for each station.

The pass–fail status for each resident was determined using a conjunctive model, which has been recommended for high-stakes pass–fail decisions, such as certification[19]. To pass the examination as a whole, candidates were required to: pass a minimum of four of eight stations; and achieve the overall examination pass score.

The agreement coefficient was calculated to assess the reliability of the pass–fail decision using Subkoviak methodology[20].

Relationship with other variables

Pass–fail status on the examination was compared with training level by means of χ^2 analysis, with the hypothesis that there is a relationship between training level and the likelihood of passing the examination.

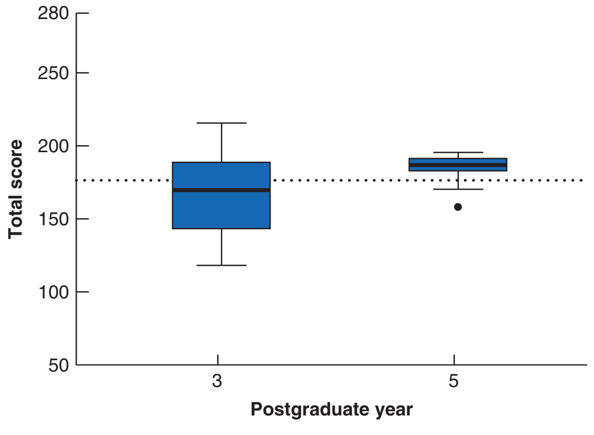
Results

Nineteen general surgery trainees participated in the study (10 PGY3, 9 PGY5). The PGY3 residents achieved a median score of 171 (range 119–217), and variability (s.d.) of 30·19. The PGY5 trainees achieved a median score of 189 (161–198), and variability of 12·16 (*Fig. 1*).

Internal structure

The interstation reliability of the examination was 0·66. Inter-rater reliability was calculated for three of the eight stations. For all three stations, the inter-rater reliability

Fig. 1 Box plot comparing overall GOSATS performance between postgraduate year 3 and 5 residents



Horizontal bars, boxes and error bars represent median, i.q.r., and range excluding outliers (symbols) respectively. An overall score of 176·8 or more represents a pass (dotted line). GOSATS, General Surgery Objective Structured Assessment of Technical Skill.

Table 1 GOSATS station and overall examination pass scores

Station	Pass score
Bleeding duodenal ulcer	23·30
Breast lumpectomy	19·90
Cholecystectomy	22·79
Haemorrhoidectomy	21·96
Small bowel resection	21·89
Inguinal hernia	21·52
Loop colostomy	21·50
Trauma laparotomy	23·92
Overall examination pass score†	176·8/280

Each station has a maximum score of 35. The maximum overall examination score is 280. GOSATS, General Surgery Objective Structured Assessment of Technical Skill.

was high at 0·92 for bleeding duodenal ulcer, 0·97 for haemorrhoidectomy, and 0·76 for trauma laparotomy and splenectomy.

Consequences evidence

The overall pass score for the examination, calculated using borderline regression standard setting methodology, was 176·8 of 280 (63·1 per cent) (*Table 1*). To pass the examination, residents were required to pass at least four of eight stations as well as to meet the pass score. Seven of nine PGY5 trainees (78 per cent) passed the examination, compared with only three of ten PGY3 trainees (30 per cent). The reliability of the pass–fail decision was high, with an agreement coefficient of 0·88.

Relationship with other variables

χ^2 analysis comparing PGY status in relation to pass–fail status demonstrated a statistically significant difference between PGY3 and PGY5 with respect to pass rates ($\chi^2 = 4{\cdot}34$, $P = 0{\cdot}037$). The odds ratio of passing the examination for PGY5 trainees compared with PGY3 trainees was 8·17 (95 per cent c.i. 1·03 to 64·94), indicating that PGY5 trainees were 8·17 times more likely to pass the examination than PGY3 trainees.

Discussion

The General Surgery Objective Structured Assessment of Technical Skill (GOSATS) examination demonstrated reliability indices that are in the range necessary for high-stakes decisions. Using a rigorous standard-setting methodology, a pass criterion for the examination was set. This study showed that graduating trainees were significantly more likely to pass the examination than PGY3 trainees, providing evidence of validity for the examination and for the set pass score. The reliability of the pass–fail decision (agreement coefficient 0·88) reflects how likely a candidate is to pass the examination on two separate administrations of the examination, assuming no additional gain in knowledge. An agreement coefficient of 0·85 or more is recommended for high-stakes decisions such as mastery examinations[17]. Guided by Messick's framework[12], this initial pilot study accrued preliminary evidence for the validity of this general surgery certification technical skills examination.

Although the current surgical training system has produced skilled surgeons for years, concerns about patient safety and the public's demand for greater accountability have provided further impetus for the development and use of more rigorous assessment tools. Simulation-based training has been demonstrated to improve technical skills, operative performance and patient outcomes[21–24]. Implementing a technical skills examination at the completion of specialty training will help ensure that new surgeons entering practice have demonstrated a minimum requisite level of competence. Although technical skill is only one domain of a competent surgeon, data suggest that the technical skills of practising surgeons can vary widely, and that surgeons with better technical skills have fewer complications, lower rates of reoperation and fewer readmissions[25].

The American Society of Colon and Rectal Surgeons has led the way in developing, implementing and evaluating a similar simulation-based examination for summative assessment of technical skills of graduating colorectal trainees: the Colorectal Objective Structured Assessment of Technical Skill (COSATS). Interestingly, the COSATS has identified technical deficiencies in individuals who passed both the oral and written colorectal board examinations, and are currently in practice[26]. This finding highlights the importance of a formal, direct and objective assessment of technical skill at the completion of specialty training.

The need to assess technical skill at the time of certification is also underscored by the complexity of the current training environment and the concept of preparedness for practice. Data suggest that limited trainee autonomy and experience, changes in the work environment, the introduction of new technologies[27] and working hour restrictions[28] have all affected trainees' preparation for independent practice[29–31]. Subspecialty fellowship programme directors have raised concerns regarding lack of readiness for advanced training. In a survey of fellowship programme directors, 30 per cent felt that incoming subspecialty trainees could not complete basic operations such as laparoscopic cholecystectomy independently, and almost one-half of programme directors (43 per cent) thought the advanced trainees could not perform 30 min of a major procedure independently in the operating room[32,33]. The development and implementation of a technical skills assessment at the time of completion of training would help to identify trainees who are not sufficiently prepared for independent operating.

Identifying individuals with technical deficiencies at the end of training offers no opportunity for remediation of skill. This raises the issue of when to assess the technical skill of surgical trainees. Although the present authors have worked to develop an examination for graduating candidates, there is ongoing discussion regarding the optimal timing for administration of the examination. Ideally, a technical skills examination should be administered before the completion of training to provide the opportunity for remediation and possible retesting. The graduating candidates in the present study were at the start of their final year, which would allow programmes enough time to identify and remediate trainees with technical deficiencies. How best to remediate these trainees is also debated and requires further discussion. Depending on the degree of difficulty, remediation may require more formal or informal training sessions, simulation-based training, dedicated on-site mentorship, proctorship during index procedures and, in some instances, repeating a clinical year. Training programmes that have implemented a remediation programme have been shown to have lower attrition rates[34]; however, little attention has been paid to how to remediate candidates who struggle with their technical skills. As this type of performance assessment becomes a reality, questions will arise as to how the failing candidate can be remediated and their skills reassessed to ensure they have reached competence.

This study has several limitations. First, it was conducted with Canadian trainees and the results may not be generalizable to other international surgical training programmes. The study should be replicated in other training programmes across the globe and modified based on their specific training objectives. Second, administration of this examination was shown to be feasible with 19 participants; however, implementation on a larger national scale to include all graduating trainees would be costly and challenging.

The study focused on the importance of technical skills assessment for certification. However, non-technical skills, such as situation awareness, decision-making and communication and leadership, are equally important attributes of a competent surgeon and require more attention when certification decisions are being made. Ideally, certification would require objective assessment of all competencies that are integral to a proficient surgeon throughout training and before entry into independent practice. Finally, although this study has built initial evidence of validity, to implement the examination in training programmes as a high-stakes examination with the potential to influence promotion and/or certification, further validity evidence will need to be accrued, with additional multicentre studies including larger sample sizes from a variety of institutions. Additionally, as the standard of assessment is predicting performance in clinical practice, further studies comparing candidates' pass–fail status on the GOSATS with their final in-training scores would help accrue validity evidence for the examination.

With further validation studies, the GOSATS may have the potential to be incorporated into competency-based certification for graduating general surgery trainees. Specialty training programmes worldwide in specialties such as vascular surgery, colorectal surgery, anaesthesia and medicine[26,35–37] have successfully incorporated some form of standardized performance-based assessment into their board certification. It is only a matter of time until the public and governing bodies require objective documentation of technical competence for all specialties.

Acknowledgements

This study was funded by a Royal College of Physicians and Surgeons of Canada Medical Education Research Grant.
Disclosure: The authors declare no conflict of interest.

References

1 Accreditation Council for Graduate Medical Education. *Milestones. Accreditation Council for Graduate Medical Education: Next Accreditation System*; 2017. https://www.acgme.org/acgmeweb/tabid/430/ProgramandInstitutionalAccreditation/NextAccreditationSystem/Milestones.aspx [accessed 11 July 2017].

2 Frank JR, Jabbour M, Fréchette D, Marks M, Valk N, Bourgeois G. *Report of the CanMEDS Phase IV Working Groups*. The Royal College of Physicians and Surgeons of Canada: Ottawa, 2005.

3 Iobst WF, Sherbino J, Cate OT, Richardson DL, Dath D, Swing SR *et al.* Competency-based medical education in postgraduate medical education. *Med Teach* 2010; **32**: 651–656.

4 Royal College of Physicians and Surgeons of Canada. *Competence by Design*. http://www.royalcollege.ca/rcsite/cbd/competence-by-design-cbd-e [accessed 11 July 2017].

5 Rubin P, Franchi-Christopher D. New edition of Tomorrow's Doctors. *Med Teach* 2002; **24**: 368–369.

6 Simpson J, Furnace J, Crosby J, Cumming AD, Evans PA, Friedman BDM *et al.* The Scottish doctor – learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. *Med Teach* 2002; **24**: 136–143.

7 Graham IS, Gleason AJ, Keogh GW, Paltridge D, Rogers IR, Walton M *et al.* Australian curriculum framework for junior doctors. *Med J Aust* 2007; **186**: S14–S19.

8 Union Européenne de Médecins Spécialistes. *Postgraduate Training*. https://www.uems.eu/areas-of-expertise/postgraduate-training; 2013 [accessed 7 July 2019].

9 Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Assessment of technical surgical skills. *Eur J Surg* 2002; **168**: 139–144.

10 Sidhu RS, Grober ED, Musselman LJ, Reznick RK. Assessing competency in surgery: where to begin? *Surgery* 2004; **135**: 6–20.

11 Gray JD. Global rating scales in residency education. *Acad Med* 1996; **71**: S55–S63.

12 Messick S. Validity. In *Educational Measurement* (3rd edn), Linn RL (ed.). American Council on Education/Macmillan series on higher education. Macmillan Publishing: New York, 1989; 13–103.

13 de Montbrun S, Louridas M, Szasz P, Harris KA, Grantcharov TP. Developing the blueprint for a general surgery technical skills certification examination: a validation study. *J Surg Educ* 2018; **75**: 344–350.

14 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative 'bench station' examination. *Am J Surg* 1997; **173**: 226–230.

15 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**: 993–997.

16 Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists *versus* global rating scales in simulation-based assessment. *Med Educ* 2015; **49**: 161–173.

17 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997; **84**: 273–278.

18 de Montbrun S, Satterthwaite L, Grantcharov TP. Setting pass scores for assessment of technical performance by surgical trainees. *Br J Surg* 2016; **103**: 300–306.

19 Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Adv Health Sci Educ Theory Pract* 1997; **2**: 201–211.

20 Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *J Educ Meas* 1988; **25**: 47–55.

21 Stefanidis D, Sevdalis N, Paige J, Zevin B, Aggarwal R, Grantcharov T *et al.*; Association for Surgical Education Simulation Committee. Simulation in surgery: what's needed next? *Ann Surg* 2015; **261**: 846–853.

22 Zendejas B, Brydges R, Hamstra SJ, Cook DA. State of the evidence on simulation-based training for laparoscopic surgery: a systematic review. *Ann Surg* 2013; **257**: 586–593.

23 McCluney A, Vassiliou M, Kaneva P, Cao J, Stanbridge DD, Feldman LS *et al.* FLS simulator performance predicts intraoperative laparoscopic skill. *Surg Endosc* 2007; **21**: 1991–1995.

24 Cox T, Seymour N, Stefanidis D. Moving the needle: simulation's impact on patient outcomes. *Surg Clin North Am* 2015; **95**: 827–838.

25 Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR *et al.*; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med* 2013; **369**: 1434–1442.

26 de Montbrun S, Roberts PL, Satterthwaite L, MacRae H. Implementing and evaluating a national certification technical skills examination: the Colorectal Objective Structured Assessment of Technical Skill. *Ann Surg* 2016; **264**: 1–6.

27 Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993; **165**: 358–361.

28 Nasca TJ, Day SH, Amis ESJ. The new recommendations on duty hours from the ACGME Task Force. *N Engl J Med* 2010; **363**: e3.

29 Friedell ML, VanderMeer TJ, Cheatham ML, Fuhrman GM, Schenarts PJ, Mellinger JD *et al.* Perceptions of graduating general surgery chief residents: are they confident in their training? *J Am Coll Surg* 2014; **218**: 695–703.

30 Coleman JJ, Esposito TJ, Rozycki GS, Feliciano DV. Early subspecialization and perceived competence in surgical training: are residents ready? *J Am Coll Surg* 2013; **216**: 764–771.

31 Napolitano LM, Savarise M, Paramo JC, Soot LC, Todd SR, Gregory J *et al.* Are general surgery residents ready to practice? A survey of the American College of Surgeons Board of Governors and Young Fellows Association. *J Am Coll Surg* 2014; **218**: 1063–1072.e1031.

32 Mattar SG, Alseidi AA, Jones DB, Jeyarajah DR, Swanstrom LL, Aye RW *et al.* General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. *Ann Surg* 2013; **258**: 440–449.

33 Guntupalli SR, Doo DW, Guy M *et al.* Preparedness of obstetrics and gynecology residents for fellowship training. *Obstet Gynecol* 2015; **126**: 559–568.

34 Schwed AC, Lee SL, Salcedo ES, Reeves ME, Inaba K, Sidwell RA *et al.* Association of general surgery resident remediation and program director attitudes with resident attrition. *JAMA Surg* 2017; **152**: 1134–1140.

35 Pandey V, Wolfe J, Liapis C, Bergqvist D; European Board of Vascular Surgery. The examination assessment of technical competence in vascular surgery. *Br J Surg* 2006; **93**: 1132–1138.

36 Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. *Anesth Analg* 2006; **102**: 853–858.

37 Medical Council of Canada. *Medical Council of Canada Qualifying Examination Part II*. https://mcc.ca/examinations/mccqe-part-ii/; 2019 [accessed 7 July 2019].