Original article

Transatlantic comparison of the competence of surgeons at the start of their professional career

M. P. Schijven[1], R. K. Reznick[3], O. Th. J. ten Cate[2], T. P. Grantcharov[3], G. Regehr[3,4], L. Satterthwaite[5], A. S. Thijssen[2] and H. M. MacRae[3]

[1]Department of Surgery, Academic Medical Center, University of Amsterdam, Amsterdam, and [2]Centre for Research and Development of Education, University Medical Centre Utrecht, Utrecht, The Netherlands, and [3]Department of Surgery, [4]Wilson Centre for Research in Education, and [5]University of Toronto Surgical Skills Centre at Mount Sinai Hospital, University of Toronto, Toronto, Canada

Correspondence to: Dr M. P. Schijven, Department of Surgery, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands (e-mail: m.p.schijven@amc.uva.nl)

Background: **Although the objective in European Union and North American surgical residency programmes is similar – to train competent surgeons – residents' working hours are different. It was hypothesized that practice-ready surgeons with more working hours would perform significantly better than those being educated within shorter working week curricula.**

Methods: **At each test site, 21 practice-ready candidate surgeons were recruited. Twenty qualified Canadian and 19 qualified Dutch surgeons served as examiners. At both sites, three validated outcome instruments assessing multiple aspects of surgical competency were used.**

Results: **No significant differences were found in performance on the integrative and cognitive examination (Comprehensive Integrative Puzzle) or the technical skills test (Objective Structured Assessment of Technical Skill; OSATS). A significant difference in outcome was observed only on the Patient Assessment and Management Examination, which focuses on skills needed to manage patients with complex problems ($P < 0{\cdot}001$). A significant interaction was observed between examiner and candidate origins for both task-specific OSATS checklist ($P = 0{\cdot}001$) and OSATS global rating scale ($P < 0{\cdot}001$) scores.**

Conclusion: **Canadian residents, serving many more working hours, perform equivalently to Dutch residents when assessed on technical skills and cognitive knowledge, but outperformed Dutch residents in skills for patient management. Secondary analyses suggested that cultural differences influence the assessment process significantly.**

Paper accepted 18 August 2009

Published online 21 January 2010 in Wiley InterScience (www.bjs.co.uk). **DOI:** 10.1002/bjs.6858

Introduction

The working hours of surgical residents have been an area of interest in recent years. With the implementation of the 80-h working week and limitations on on-call duty, working hours for surgical residents in North America have been reduced substantially. This is even more marked in the European Union (EU), where a maximum working week of 48 h is mandated. The impact of working hour restrictions on the quality of surgical training has been debated, but there is little direct evidence of the effect of decreased working hours on surgical competency[1–12]. Studies suggest an improved satisfaction with work–life balance, but at a perceived cost of potentially less complete training.

Although the objective in the EU and North American surgical residency programmes is similar – to train competent surgeons – the actual time spent in active patient care during surgical residency is not. This study investigated the impact of working hour restrictions on the competence of surgeons at completion of training. Cohorts of practice-ready surgeons from Canada and the Netherlands were compared on three validated tests of clinical competence, and differences in performance were assessed. It was hypothesized that practice-ready surgeons exposed to

The Editors are satisfied that all authors have contributed significantly to this publication

a longer working hour training programme would perform significantly better than practice-ready surgeons exposed to a shorter working hour training. A secondary hypothesis was that examiner country of origin might have an impact on how the candidate was evaluated.

Methods

Three instruments were chosen for a multifaceted comparison of clinical competency. The Comprehensive Integrative Puzzle (CIP), developed by Ber[13], was chosen to assess knowledge and analytical clinical reasoning. For assessment of the competence of senior surgical residents in patient management, communication skills and problem solving, the Patient Assessment and Management Examination (PAME) was used[14]. Finally, the Objective Structured Assessment of Technical Skill (OSATS) was used to assess technical skill[15]. All tests were administered within 1 day at each of the Canadian and Dutch testing sites. Examinations were similar in content, and administered in the Dutch language in the Netherlands and in the English language in Canada. Translation of tests from English to Dutch was checked by native speakers. This research focused on the end-result of residency training.

Assessment instruments

Comprehensive Integrative Puzzle

To assess knowledge and clinical reasoning, the CIP, an extended matching test, was used. Participants were asked to complete matrices for five surgical domains (vascular surgery, surgical oncology, gastrointestinal surgery, paediatric surgery and trauma surgery). The first column of each matrix displays five 'cases' of common surgical syndromes or diseases, with the cases set against seven answering columns reflecting diagnostic or therapeutic categories. The CIP score reflects scoring from a possible 175 items. Test time was fixed at 90 min.

Patient Assessment and Management Examination

The PAME is a standardized patient-based examination designed for summative assessment of senior surgical residents in which a sequence of events in clinical practice is simulated through six patient encounters. Each 25-min station includes a training script for the standardized patient, referral letters, imaging studies, endoscopy images and laboratory results. After each simulated patient encounter, a short, structured, oral examination is administered. Examiners rate residents' performance with global rating scales[14]. The PAME has been advocated as a reliable and valid tool to address evaluation of several Accreditation Council for Graduate Medical Education competencies that are otherwise difficult to assess[16]. Station content was chosen by the core faculty from both sites, to ensure face and content validity. Stations selected were: 'hereditary non-polyposis colorectal carcinoma', 'breast lesion in a pregnant female', 'mother of an infant with pyloric stenosis', 'adrenal incidentaloma', 'right upper quadrant pain' and 'Zenker's diverticulum'.

Objective Structured Assessment of Technical Skill

The OSATS consisted of eight stations where residents performed operative procedures on bench models. The following stations were chosen: laparoscopic suturing, vascular anastomosis, tracheostomy, ileostomy, J-tube insertion, rectal anastomosis, inguinal hernia and laparoscopic cholecystectomy. Checklist and global rating scale scores were used to assess performance. Test time was fixed at 12 min per station plus turnover time for a total test time of 2 h.

Half of the candidates took the PAME first, and half took the OSATS first at each test site (allocated randomly), and then switched to the other examination. All residents completed the CIP simultaneously, but individually, at the end of the day.

Candidates

Assuming that a difference of 10 per cent or more (mean difference calculated over all tests) was of relevance ($d = 0{\cdot}10$), with an anticipated standard deviation of 10 per cent ($s = 0{\cdot}1$), a power ($1 - \beta$) of 0·9 and $\alpha = 0{\cdot}05$, a sample size of 21 residents per assessment group was obtained.

At each site, 21 candidates who were within 6 months of certification as a surgeon (before or after certification) were recruited to participate in the study, Canadian residents at the end of year 5 and Dutch residents at the end of year 6. At the Canadian test site, the majority of candidates were scheduled to complete their training within 2 months. Three of the candidates had completed general surgical training 10 months previously, and were now in fellowship training; they were thus more experienced. Candidates were representative of multiple surgical training sites across both countries. All candidates were novice to CIP, PAME and OSATS testing.

Estimated working hours

A study to estimate the actual working hours spent by residents at each test site preceded the main study. During a 2-week period, a cross-sectional cohort of residents who

did not participate in the study, but who were working in the test site hospitals, filled in a working hour registration list. The mean number of hours a Dutch resident worked during one working week of 5 days, including evening/night shifts in the Utrecht Medical Centre in the Netherlands, was 55 h (19 residents, 100 per cent response rate). The mean number of hours for a similar sample of residents at the University of Toronto in Canada was 84 h per working week.

Faculty

In Canada, 20 Canadian surgeons assessed the candidates on either OSATS or PAME stations, and five Dutch surgeons rated candidates on either OSATS (three) or PAME (two). In the Netherlands, 19 Dutch surgeons assessed the candidates on either OSATS or PAME, and six Canadian surgeons assessed on OSATS.

Statistical analysis

Four primary outcome scores were calculated: CIP score, PAME score, OSATS checklist (OSATS-C) score and OSATS global rating scale (OSATS-G) score. The reliability of each score was assessed using Cronbach's α. For the OSATS, interrater reliability was determined on stations where candidates were assessed by two examiners simultaneously. To address the primary research question (a comparison of performance between the two sites), the Canadian and Dutch means were compared for each outcome measure using Student's *t* test.

Results

Reliability of outcome measures

Table 1 shows the internal consistency analysis of the four outcome measures evaluated. For OSATS, there were several stations where only one examiner (Canadian or Dutch) produced complete data for all candidates. When there was no second examiner at a station, or an incomplete data set was generated by an examiner, the station was eliminated from the relevant analysis.

For three OSATS stations, at which the same raters evaluated as pairs at both the Canadian and Dutch sites, interrater reliability (Pearson's correlation coefficient) was also calculated for the pairs of examiners observing the same candidates at these stations. The individual station interrater correlations for OSATS-C scores at these stations were 0·65, 0·43 and 0·89, giving an interrater reliability of 0·66. For OSATS-G scores, the interrater correlations for the same stations were 0·77, 0·42 and 0·82 respectively, giving an interrater reliability of 0·67.

Correlation coefficients between the four primary outcome measures are presented in *Table 2*. The high correlation between the two OSATS measures (0·80) is not surprising given that they were generated by the same raters observing the same performances. The low correlations between the other measures is in part a function of the moderate reliabilities of the respective measures, as presented in *Table 1*, but also provides some evidence of divergent validity, that is the complementary aspect of the various primary outcome measures.

Table 1 Cronbach's α for the four primary outcome measures

	All candidates ($n = 42$)	Canadian candidates ($n = 21$)	Dutch candidates ($n = 21$)
CIP*	0·73 (175)	0·64 (175)	0·81 (175)
PAME			
Canadian examiners		0·71 (6)	
Dutch examiners			0·50 (6)
Combined sites*	0·72 (6)		
OSATS-C			
Canadian examiners		0·25 (8)	0·35 (5)
Dutch examiners		0·12 (3)	0·39 (6)
Combined scores*	0·20 (8)	0·23 (8)	0·29 (8)
OSATS-G			
Canadian examiners		0·53 (8)	0·45 (5)
Dutch examiners		0·24 (3)	0·43 (6)
Combined scores*	0·48 (8)	0·53 (8)	0·53 (8)

Values in parentheses are the number of items/stations represented in the calculation. *Outcome measure used for primary comparisons between sites. CIP, Comprehensive Integrative Puzzle; PAME, Patient Assessment and Management Examination; OSATS-C, Objective Structured Assessment of Technical Skill checklist; OSATS-G, Objective Structured Assessment of Technical Skill global rating scale.

Table 2 Between-measurements correlation coefficients (Pearsons' product moment) of the four primary outcome measurements on both test sites combined

	CIP	PAME	OSATS-C	OSATS-G
CIP	n.a.	0·03	0·21	0·19
PAME	0·03	n.a.	0·20	0·08
OSATS-C	0·21	0·20	n.a.	0·80
OSATS-G	0·19	0·08	0·80	n.a.

CIP, Comprehensive Integrative Puzzle; PAME, Patient Assessment and Management Examination; OSATS-C, Objective Structured Assessment of Technical Skill checklist; OSATS-G, Objective Structured Assessment of Technical Skill global rating scale; n.a., not applicable.

Table 3 Student's *t* test (two tailed) for Canadian and Dutch residents on the four primary outcome measures

	Canadian candidates	Dutch candidates	t_{40}	*P*
CIP	0·52(0·05)	0·53(0·07)	0·18	0·856
PAME	0·85(0·06)	0·79(0·05)	3·90	< 0·001
OSATS-C	0·78(0·06)	0·75(0·06)	1·33	0·192
OSATS-G	0·75(0·06)	0·74(0·07)	0·66	0·515

Values are mean(s.d.). CIP, Comprehensive Integrative Puzzle; PAME, Patient Assessment and Management Examination; OSATS-C, Objective Structured Assessment of Technical Skill checklist; OSATS-G, Objective Structured Assessment of Technical Skill global rating scale.

Primary analyses: comparison of residents' performance at the two sites

Student's *t* tests (independent samples) were used to compare the Canadian and Dutch candidate scores (*Table 3*). No significant differences in measures of knowledge (CIP) or technical skill/knowledge (OSATS) were observed. Significant differences were observed in the patient assessment and management examination (PAME).

Secondary analyses: cultural interpretations for quality of performance

The secondary set of hypotheses related to the potential for 'cultural' differences in what was to be considered an excellent performance among the Canadian and Dutch examiners. Both Canadian and Dutch examiners were used at selected stations to assess the Canadian and Dutch candidates similarly. To assess for potential cultural effects, a set of two-way, mixed design, ANOVA was performed for OSATS-C and OSATS-G. Only stations where the examiner observed at both sites were taken into the equation, as subanalysis on the influence of faculty presence on scoring (taking into account all available data, data sets of only those examiners examining in two countries, and only data sets of examiners present in two countries on the same stations) showed consistent patterns even in this most conservative analysis. For both the checklist and global rating scores, Canadian examiners scored Canadian candidates higher than Dutch candidates, and Dutch examiners scored Dutch candidates higher than Canadian candidates (*Fig. 1*). Consistent with these patterns, ANOVA demonstrated a significant interaction between examiner origin and candidate origin for both the checklist ($F_{1,40} = 13{\cdot}62$, $P = 0{\cdot}001$) and global rating ($F_{1,40} = 22{\cdot}61$, $P < 0{\cdot}001$) scores.

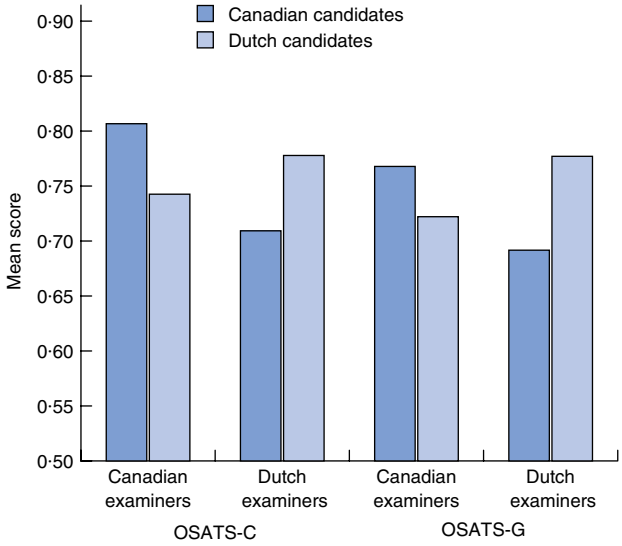

Fig. 1 Comparison of Objective Structured Assessment of Technical Skills checklist (OSATS-C) and global rating scale (OSATS-G) scores for Dutch and Canadian candidates assessed by Dutch and Canadian examiners (two-way mixed ANOVA design)

Discussion

Working hour restrictions have been implemented in many jurisdictions owing to concerns around patient safety and resident well-being[7,17–20]. Although it is recognized that restriction of working hours must have some influence on surgical training[12], the impact of such a reduction has not been assessed previously. To the authors' knowledge, no previous study has attempted directly to examine the impact of these restrictions in terms of outcome of surgical competency.

The implications of decreased working hours have been studied in the USA[21–24] and the Netherlands[25]. In Canada, residents' working hours are restricted to in-house calls for one night in four. Thus, the 'least hour working week' scenario in Canada is 72 h per week, although working 80 h per week is reported to be fairly common, as was found in the authors' preliminary analysis of working hours. In the Netherlands, residents' working hours are limited to 48 h per week, including being on call, following the European Working Hours Directive. If the average surgical resident has 4 weeks' annual leave, the difference in number of hours between North American residents at 80 h per week and Dutch residents at 48 h per week is more than 7000 h. This translates, in theory, to a difference of over 2 years' full-time training. It has been suggested, in both Canada and the Netherlands, that the legislative working week is not necessarily honoured[7,26,27]. In an analysis of moonlighting

experience that preceded this study for working hours, both academic centres showed a similar proportion of 15 per cent more working hours per week than the maximum allowed. Although hours spent do not necessarily equate with increased educational opportunities, it seems reasonable to assume that competency outcome, the key question for this study, may suffer as a result of fewer hours in the clinical environment.

At present, competency-based frameworks are implemented to structure surgical training programmes in many EU and North American surgical residency programmes. These frameworks help to ensure that programmes are oriented towards their primary goal, the training of competent surgical practitioners[17,28–30]. Nevertheless, despite a common endpoint, training programmes in Europe and North America have large differences in the number of working hours.

In this study, three outcome instruments, the CIP, the OSATS and the PAME, were used to assess a variety of competencies needed for practice-ready surgical performance. Assessment groups were standardized to be near practice-ready, with similar average distances to this point in both groups. There were no significant differences between Canadian and Dutch candidates on either the CIP or the OSATS. Both tests were shown to be reliable, and were previously validated in their ability to test integrative cognitive and clinical knowledge and technical surgical skill respectively[31–38]. This suggests that reduced hours in the hospital do not lead to a measurable decrement in either one of these two domains of performance. This finding is consistent with a previous report, which suggested that operative volumes do not suffer substantially when working hour restrictions are implemented[8]. In the Netherlands, a cohort study was unable to demonstrate a decrease in the number of patients operated on by trainees for the period 2000–2005[7].

Despite implementation of working hour restrictions, Guicherit[25] found stable rates of operative experience for trainees in the Netherlands between 1990 and 2000. Perhaps experience in the operating room is the clinical activity least likely to be affected by working hour restrictions, as residents tend to be motivated to ensure they have adequate operating exposure. Whether a reduction in working hours per week implies less operative procedural exposure or not, it must be noted that that 'operative exposure' is not synonymous with 'surgical competency'. Even if shortening the working week has not led to less operative procedural exposure, it must come at the expense of something else. Previously, it has been shown in the Netherlands that this was at the expense of residents spending less time on the wards and in the clinic, that is, having less time for interacting and communicating with patients[7].

The PAME was the examination of choice to assess handling complex and integrated patient assessment and management skills. As the examination was administered in candidates' native language, no Canadian scores could be obtained on the Dutch test site for PAME, and only a small Dutch examiner data set (one assessment track of six stations by two examiners) was available for PAME on the Canadian test site. Therefore, only Canadian rater data assessing Canadian candidates and only Dutch rater data assessing Dutch candidates were used in this analysis. From these examination data, a significant difference between the two groups of residents was found. The magnitude of the difference was similar to that seen between fourth- and fifth-year residents at the University of Toronto in a previous study[14]. The PAME focuses on many aspects of performance that are best attained through interactions with patients around complex surgical problems, and these abilities take time to develop. As mentioned previously, shortened working weeks may come at the expense of surgical residents spending less time on the wards and the clinic, with the implication that residents have potentially less time to master the complicated interactions required for good clinical practice[7].

A critical finding in this study was that, on the stations where it could be assessed, Dutch and Canadian examiners evaluated candidates differently, with examiners favouring candidates from their own country. This pattern was consistent and seen throughout multiple analyses, on full international data-matched observer stations and on partial international matched assessment stations. This was true for both the PAME and the OSATS. Global ratings, as used in the OSATS and PAME, rely on expert analysis of performance, with raters developing a holistic impression of how well the candidate is functioning. In the present study, expert ratings seemed to be somewhat country specific, with examiners preferring the approach of candidates from their own setting. Indeed, Dana[39] reported that clinician bias occurs during test/method administration, based on the clinician's own cultural and/or racial identity. Bias among clinicians is thought to be typically inadvertent, denied, and often below the threshold of awareness. It is thus of utmost important for examining clinicians to eliminate eisegesis (intrusion of examiner personality) in the interpretation of test circumstances and performance.

In conclusion, in this first transatlantic comparison of the skills of surgical residents at the conclusion of their training, various surgical competencies were assessed. Residents from the Netherlands with fewer working hours

performed similarly to Canadian residents in technical surgical skill, as measured with the OSATS, and in the area of knowledge and clinical reasoning, as measured by the CIP. On PAME, an examination focusing on complex skills needed for management of patients with complex problem scenarios, a significant difference in outcome was observed, favouring Canadian trainees with more working hours. Significant differences were seen in the use of expert-based evaluations by examiners from different countries watching the same performance. As this type of research is challenging, the use of digital recordings for cultural assessments may prove to be useful in future research.

As the world becomes increasingly globalized, with increasing access to services between countries, it is inevitable that there will at some point be a demand for standardized curricula and assessment procedures, resulting in a global standard of credentialing. The results of this study suggest that, if this eventuates, further work will be needed to objectify the assessment process in order to mitigate the cultural differences that may influence it. The authors believe that future studies aimed at investigating differences in surgical competency between trainees from different practice environments would benefit from using raters who are independent of the particular practice environments involved. They feel strongly that surgical residents should be assessed on their surgical competencies periodically by unrelated surgical examiners (preferably not from their own hospital). There should be explicit training procedures for examiners, in order to minimize factors that contribute to systematic bias.

Acknowledgements

This study would not have been possible without the enthusiasm and help of many individuals. First, the authors wish to thank all candidates, faculty members, simulative patients, and assistants on both assessment days. Porcine material was given free of charge by VION Food Group, The Netherlands. G.R. is supported as the Richard and Elizabeth Currie Chair in Health Professions Education Research. Funding was generously provided by both university hospitals. The authors declare no conflict of interest.

References

1 Barden CB, Specht MC, McCarter MD, Daly JM, Fahey TJ III. Effects of limited work hours on surgical training. *J Am Coll Surg* 2002; **195**: 531–538.
2 Steinbrook R. The debate over residents' work hours. *N Engl J Med* 2002; **347**: 1296–1302.
3 Chung RS. How much time do surgical residents need to learn operative surgery? *Am J Surg* 2005; **190**: 351–353.
4 Schenarts PJ, Anderson Schenarts KD, Rotondo MF. Myths and realities of the 80-hour work week. Review. *Curr Surg* 2006; **63**: 269–274.
5 Lin GA, Beck DC, Garbutt JM. Residents' perceptions of the effects of work hour limitations at a large teaching hospital. *Acad Med* 2006; **81**: 63–67.
6 Irani JL, Mello MM, Ashley SW, Whang EE, Zinner MJ, Breen E. Surgical residents' perceptions of the effects of the ACGME duty hour requirements 1 year after implementation. *Surgery* 2005; **138**: 246–253.
7 Wijnhoven BP, Watson DI, van den Ende ED. Current status and future perspective of general surgical trainees in the Netherlands. *World J Surg* 2008; **32**: 119–124.
8 Spencer AU, Teitelbaum DH. Impact of work-hour restrictions on residents' operative volume on a subspecialty surgical service. *J Am Coll Surg* 2005; **200**: 670–676.
9 McElearney ST, Saalwachter AR, Hedrick TL, Pruett TL, Sanfey HA, Sawyer RG. Effect of the 80-hour work week on cases performed by general surgery residents. *Am Surg* 2005; **71**: 552–555.
10 Kaafarani HM, Itani KM, Petersen LA, Thornby J, Berger DH. Does resident hours reduction have an impact on surgical outcomes? *J Surg Res* 2005; **126**: 167–171.
11 Durkin ET, McDonald R, Munoz A, Mahvi D. The impact of work hour restrictions on surgical resident education. *J Surg Educ* 2008; **65**: 54–60.
12 Glomsaker TB, Søreide K. Surgical training and working time restriction. *Br J Surg* 2009; **96**: 329–330.
13 Ber R. The CIP (comprehensive integrative puzzle) assessment method. *Med Teach* 2003; **25**: 171–176.
14 MacRae H, Cohen R, Regehr G, Reznick R, Burnstein M. A new assessment tool: the patient assessment and management examination. *Surgery* 1997; **122**: 335–343.
15 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.
16 Dunnington GL, Williams RG. Addressing the new competencies for residents' surgical training. *Acad Med* 2003; **78**: 14–21.
17 Accreditation Council for Graduate Medical Education. *ACGME-Common Program Requirements for Resident Duty Hours* (1 June 2003; Updated 1 July 2007). http://www.acgme.org/acWebsite/dutyhours/dh_index.asp [accessed 8 February 2009].
18 Accreditation Council for Graduate Medical Education. *A Summary of Achievement of the ACGME's Approach to Limit Resident Duty Hours, Academic Year 2007–08*. http://www.acgme.org/acWebsite/dutyhours/dh_index.asp [accessed 7 November 2009].
19 Royal College of Surgeons of England. *Surgical Training Seriously Compromised By European Working Time Directive*. 2005; http://www.rcseng.ac.uk/media/medianews/Surgicaltrainingcompromisedbyworkingtimedirective [accessed 7 November 2009].

20 Benes V. The European Working Time Directive and the effects on training of surgical specialists (doctors in training): a position paper of the surgical disciplines of the countries of the EU. *Acta Neurochir* 2006; **148**: 1227–1233.
21 Mathis BR, Diers T, Hornung R, Ho M, Rouan GW. Implementing duty-hour restrictions without diminishing patient care or education: can it be done? *Acad Med* 2006; **81**: 68–75.
22 Kort KC, Pavone LA, Jensen E, Haque E, Newman N, Kittur D. Resident perceptions of the impact of work-hour restrictions on health care delivery and surgical education: time for transformational change. *Surgery* 2004; **136**: 861–871.
23 Coverdill JE, Adrales GL, Finlay W, Mellinger JD, Anderson KD, Bonnell BW *et al*. How surgical faculty and residents assess the first year of the Accreditation Council for Graduate Medical Education duty-hour restrictions: results of a multi-institutional study. *Am J Surg* 2006; **191**: 11–16.
24 Lockley SW, Cronin JW, Evans EE, Cade BE, Lee CJ, Landrigan CP *et al*.; Harvard Work Hours Health and Safety Group. Effect of reducing interns' weekly work hours on sleep and attentional failures. *N Engl J Med* 2004; **351**: 1829–1837.
25 Guicherit OR. Operative volume of surgeon trainees registered in the period 1990–2000. *Ned Tijdschr Heelk* 2002; **11**: 13–17.
26 Landrigan CP, Barger LK, Cade BE, Ayas NT, Czeisler CA. Interns' compliance with accreditation council for graduate medical education work-hour limits. *JAMA* 2006; **296**: 1063–1070.
27 Woodrow SI, Park J, Murray BJ, Wang C, Bernstein M, Reznick RK *et al*. Differences in the perceived impact of sleep deprivation among surgical and non-surgical residents. *Med Educ* 2008; **42**: 459–467.
28 Royal College of Physicians and Surgeons of Canada. *The CanMEDS Physician Competency Framework. Better Standards, Better Physicians, Better Care.* 2007; http://rcpsc.medical.org/canmeds/index.php [accessed 7 November 2009].
29 Frank JR, Langer B. Collaboration, communication, management, and advocacy: teaching surgeons new skills through the CanMEDS Project. *World J Surg* 2003; **27**: 972–978.
30 Hamming J, Borel Rinkes I, Heineman E. *Scherp (Structuur Curriculum Heelkunde voor Reflectieve Professionals). Opleidingsplan Heelkunde*. 2008; http://knmg.artsennet.nl/web/file?uuid=e9ca357d-dab2-4a93-9c8e-549a77e2c647&owner=db8aa948-9eba-4439-afc5-dedd737f7886 [accessed 7 November 2009].
31 Groothoff JW, Frenkel J, Tytgat GAM, Vreede WB, Bosman DK, ten Cate OT. Growth of analytical thinking skills over time as measured with the MATCH test. *Med Educ* 2008; **42**: 1037–1043.
32 MacRae H, Regehr G, Leadbetter W, Reznick RK. A comprehensive examination for senior surgical residents. *Am J Surg* 2000; **179**: 190–193.
33 Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *Am J Obstet Gynecol* 2006; **195**: 617–621.
34 Bodle JF, Kaufmann SJ, Bisson D, Nathanson B, Binney DM. Value and face validity of objective structured assessment of technical skills (OSATS) for work based assessment of surgical skills in obstetrics and gynaecology. *Med Teach* 2008; **30**: 212–216.
35 Reznick R, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinée D. Process-rating forms *versus* task-specific checklists in an OSCE for medical licensure. Medical Council of Canada. *Acad Med* 1998; **73**(Suppl 10): S97–S99.
36 Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993; **165**: 358–361.
37 Reznick RK, MacRae H. Teaching surgical skills – changes in the wind. *N Engl J Med* 2006; **355**: 2664–2669.
38 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.
39 Dana RH. *Multicultural Assessment Principles, Applications, and Examples*. Lawrence Erlbaum Associates: Mahwah, 2005.