The Royal College of Surgeons of England

ORIGINAL ARTICLE

Ann R Coll Surg Engl 2005; **87**: 242–7
doi 10.1308/003588405X51100

Developing assessments of surgical skills for the GMC Performance Procedures

JONATHAN D BEARD[1], BRIAN C JOLLY[2], LESLEY J SOUTHGATE[5], DAVID I NEWBLE[4], EG THOMAS[5], JOHN ROCHESTER[6]

[1]Programme Director for Higher Surgical Training, Sheffield Teaching Hospitals, Sheffield, UK
[2]Department of Medical Education, Monash University, Australia
[3]Performance Assessment Implementation Group, UK General Medical Council, London, UK
[4]Department of Medical Education, University of Sheffield, Sheffield, UK
[5]Surgical Tutor at The Royal College of Surgeons of England, Sheffield Teaching Hospitals, Sheffield, UK
[6]Department of Surgery, Rotherham General Hospital, Rotherham, UK

ABSTRACT

INTRODUCTION The objectives were to: (i) establish how 'typical' consultant surgeons perform on 'generic' (non-specialist) surgical simulations before their use in the General Medical Council's Performance Procedures (PPs); (ii) measure any differences in performance between specialties; and (iii) compare the performance of group of surgeons in the PPs with the 'typical' group.

VOLUNTEERS AND METHODS Seventy-four consultant volunteers in gastrointestinal surgery (n = 21), vascular surgery (n = 11), urology (n = 10), orthopaedics (n = 15), cardiothoracic surgery (n = 10) and plastic surgery (n = 7), plus 9 surgeons undertaking phase 2 of the PPs undertook 7 simple simulations in the skills laboratory. The scores of the volunteers were analysed by simulation and specialty using ANOVA. The scores of the volunteers were then compared with the scores of the surgeons in the PPs.

RESULTS There were significant differences between simulations, but most volunteers achieved scores of 75–100%. There was a significant simulation by specialty interaction indicating that the scores of some specialties differed on some simulations. The scores of the group of surgeons in the PPs were significantly lower than the reference group for most simulations.

CONCLUSIONS Simple simulations can be used to assess the basic technical skills of consultant surgeons. The simulation by specialty interaction suggests that whilst some skills may be generic, others are not. The lower scores of the surgeons in the PPs suggest that these tests possess criterion validity, *i.e.* they may help to determine when poor performance is due to lack of technical competence.

KEYWORDS
Surgery – Skills – Simulations – Competence – Assessment

CORRESPONDENCE TO
Mr J D Beard, Programme Director, Sheffield Vascular Institute, Northern General Hospital, Sheffield S5 7AU, UK
T: +44 (0)114 271 5534; F: +44 (0)114 271 4747; E: Jonathan.D.Beard@sth.nhs.uk

The introduction of the General Medical Council's (GMC's) Performance Procedures (PPs) requires doctors whose registration has been called into question to undergo an assessment of performance.[1] All doctors in the programme are assessed within a generic framework derived from the GMC's guidance *Good Medical Practice*.[2] Phase 1 comprises a peer review of performance in the workplace by two medical assessors and one lay assessor. Phase 2 includes standardised objective tests of knowledge, communication and technical skills that assess competence in a testing centre. Competence is a necessary prerequisite for performance and these tests are designed to clarify whether the basis of poor performance is incompetence or other factors such as illness, stress or environment.[3] The output from these tests of competence is one small part of a much wider body of evidence (triangulation) that is required for such a high-stakes assessment.

The primary aim of this study was to establish how 'typical' consultant surgeons perform on 'generic' (non-specialist) surgical simulations before their use in the GMC's PPs. The secondary aims were to measure any inter-specialty differences between these 'generic' simulations to see whether

Table 1 The seven generic simulations covering the skills expected of all surgeons

Scrubbing up, then donning a gown and gloves to assess aseptic technique
Preparing a patient (model) for an operation including the necessary safety checks, positioning on the operating table, diathermy placement, *etc*. The operation depends upon the specialty, *e.g.* total hip replacement for orthopaedic surgeons
Hand and instrument knotting on a knotting rig, including knotting at depth
Skin incision and suturing on a skin pad (interrupted suturing with an instrument and subcuticular hand suturing)
Ligation of vessels (division between haemostats and ligation in continuity) using porcine small bowel mesentery
Dissection of tissue assessed by excision of a lymph node from porcine small bowel mesentery
Hand–eye co-ordination using an endoscopic trainer to assess the transfer of objects from one instrument to another, followed by excision of an area marked out on a rubber glove and application of clips to the pedicle

they represented a fair test for all surgeons, regardless of specialty, and to compare the scores achieved by a group of surgeons in the PPs.

Study Group and Methods

All operations can be broken down into a series of core skills. The generic simulations selected for the GMC PPs were mostly derived from the Basic Surgical Skills Course and the Specialist Registrar Skills Course in general surgery developed by The Royal College of Surgeons of England.[4] These seven simulations cover the technical skills that might be expected of all surgeons (Table 1).

The technical skills' assessments for the volunteers used a format similar to an Objective Structured Clinical Examination (OSCE). Precise instructions were provided for each simulation, with 15 min allocated per simulation. Each simulation was scored using a task-specific checklist of 15–20 items, weighted if necessary, to give a maximum score of 20. These check lists were derived from those developed by Winckel *et al.*[5] for their Objective Structured Assessment of Technical Skills (OSATS). The simulations were validated against performance in the operating theatre on a group of 33 surgical trainees.[6]

Seventy-four consultant volunteers in gastrointestinal surgery (n = 21), vascular surgery (n = 11), urology (n = 10), orthopaedics (n = 15), cardiothoracic surgery (n = 10) and plastic surgery (n = 7) were assessed. Volunteers included surgical assessors for the PPs, local colleagues from South Yorkshire and others with an interest in the assessment of surgical competence. Each simulation was marked by a separate examiner who remained at that station.

Nine surgeons who had not performed well above the standard for registration on the Phase 1 PPs' assessments were assessed in a similar way. There were a few differences in the design of the PPs' assessments due to the way in which the performance rules are written, largely in the interests of fairness to the doctors. First, no time limit was set to reduce stress. Second, two assessors marked each doctor on all simulations. Hence, in the PPs, the number of judgements made on each surgeon was 14, whereas in the volunteer study it was 7. This is likely to make the PPs' scores more precise (reproducible) than the scores generated for the volunteers. We would have liked to have replicated the PPs design for the volunteers but this was impractical for logistical and financial reasons.

Statistical analysis

The median, interquartile and range of the scores for each simulation and each group were calculated, together with the mean and 95% confidence intervals (CIs) which were required for subsequent analysis. A two-way analysis of variance (ANOVA) was used to analyse the relative contribution of differences between specialties, simulations, and their interactions using the GENOVA program designed by Crick and Brennan.[7] This produced estimates of specialty and station by specialty mean squares and associated F-tests. The differences between the specialties on each simulation were then compared in more detail using the Student-Newman-Keuls (SNK) method. SNK identifies homogenous sub-groups of specialties with respect to performance on each measure. Further analysis of all possible specialty pairs, using the Bonferroni technique, which corrects significance levels for multiple comparisons, was also performed.

Results

Group demographics

Four of the 74 volunteers and none of the 9 surgeons in the PPs were women. The median age of the two groups was 41

years (range, 33–58 years) and 53 years (range, 40–67 years), respectively.

Simulation scores

There were small differences in the median scores and ranges for each simulation varying from 20 (range, 13–20) for dissection to 16 (range, 8–20) for scrubbing (Fig. 1). The median scores and ranges for simulations by specialties are shown in Table 2.

Differences between specialties

ANOVA (Table 3) showed significant differences between simulations (F = 3.85; df (6, 33); $P = 0.005$), which reflects inherent differences in simulation difficulty, and a specialty by simulation interaction (F = 2.21; df (30, 408); $P = 0.0001$). This reflects differences in scores between specialties on some simulations (Fig. 2). Subsequent SNK comparisons between specialties confirmed that orthopaedic and plastic surgeons scored lower than all other groups in suturing and tissue dissection and vessel ligation. Pairwise comparisons using Bonferroni confirmed these findings. There were 3 extreme outliers with poor hand–eye co-ordination scores. None of these three surgeons undertook any endoscopic surgery. The one extreme outlier in the tissue dissection simulation was an orthopaedic surgeon who did not undertake any soft-tissue surgery.

Comparison with surgeons in the PPs

The mean scores of the surgeons in the PPs were significantly lower for all simulations than the volunteer group for each simulation ($P = 0.0001$) except for hand–eye co-ordination ($P = 0.004$), scrubbing and patient preparation (not significant). Their overall scores were also significantly worse ($P = 0.0001$) and their interquartile range did not overlap with any other specialty group (Fig. 2).

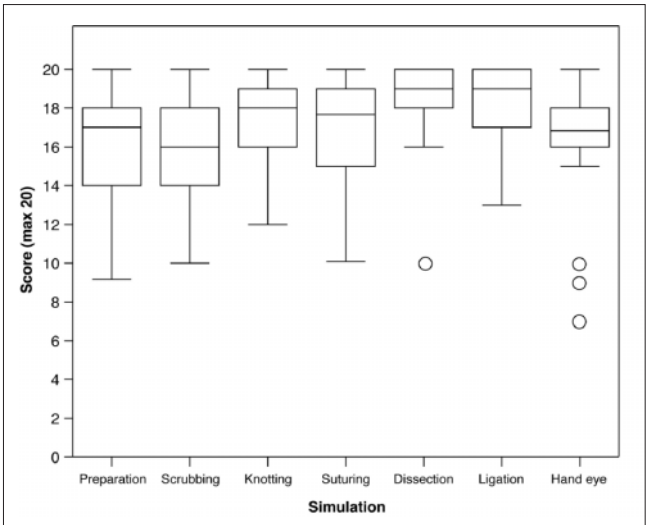

Figure 1. Box and whisker plot (median = bar, interquartile range = box and whiskers = range) of the scores for the 74 volunteers by simulation (circles denote extreme outliers).

Discussion

There are few studies comparing the results obtained in the skills laboratory with those in the operating theatre. Parallel examinations of technical skill, one using live animals and one using simulations, have been developed by the Department of Surgery, University of Toronto, Ontario, Canada.[5] Twenty surgical residents each took both formats, which were assessed using an OSATS format. The correlations between live and bench scores were high (0.69–0.72), and the mean inter-rater reliability between simulations ranged from 0.64–0.72. Paisley *et al.*[8] found non-significant or weak correlations between the technical skills of basic surgical trainees assessed in the skills laboratory compared with overall ratings by their consultant trainer. However, more recent work has found that the simulations used in this study correlated highly with direct and video observation of performance on specific operations.[6]

The assessment of technical skills in this study uses task-specific checklists. Martin *et al.*[9] have found better reliability using global rating scales; however, these seem difficult to use for more simple models, as many of the elements of a global rating become inapplicable. Global rating scales appear better suited to more the complex specialty simulations, or for direct observation in the operating theatre.

The volunteers were not randomly recruited and the numbers are small. However, the difficulty and costs of

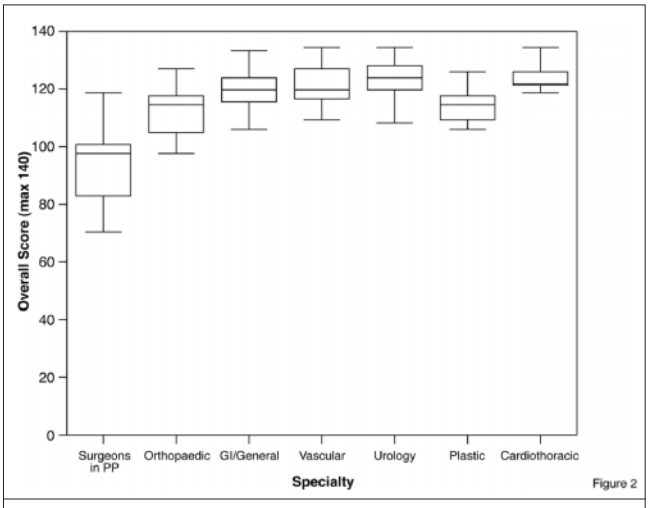

Figure 2. Box and whisker plot of the overall scores for the volunteers by specialty and the 9 surgeons in the Performance Procedures.

Table 2 Mean and median scores, 95% CI and ranges for each specialty, all volunteers and surgeons in the PPs for each of the seven simulations

Simulation	Specialty	Mean	95% CI		Median	Range	
			Lower	Upper		Min	Max
Preparation							
	Orthopaedics	15.1	13.2	17.0	16.0	7	19
	GI/general	14.3	12.5	16.1	15.0	7	20
	Vascular	16.4	14.2	18.6	17.0	9	20
	Urology	17.2	15.2	19.2	18.0	10	20
	Plastic	16.9	14.4	19.3	17.0	12	20
	Cardiothoracic	17.0	15.5	18.5	17.0	13	20
	All volunteers	15.8	15.0	16.6	17.0	7	20
	Surgeons in PPs	14.9	12.1	17.6	16.0	7	19
Scrubbing							
	Orthopaedics	16.4	14.9	17.9	16.0	8	20
	GI/general	14.5	12.9	16.2	14.0	8	20
	Vascular	15.4	13.4	17.4	16.0	10	20
	Urology	16.4	14.0	18.8	17.0	11	20
	Plastic	17.0	15.8	18.2	17.0	15	19
	Cardiothoracic	17.5	16.4	18.6	18.0	15	20
	All volunteers	15.9	15.2	16.6	16.0	8	20
	Surgeons in PPs	13.9	11.1	16.8	14.5	7	18
Knotting							
	Orthopaedics	17.2	15.7	18.7	18.0	10	20
	GI/general	17.2	16.0	18.4	18.0	11	20
	Vascular	17.2	15.5	18.9	18.0	12	20
	Urology	17.0	15.7	18.3	17.0	14	20
	Plastic	16.4	14.3	18.6	15.0	14	20
	Cardiothoracic	18.4	17.4	19.3	19.0	15	20
	All volunteers	17.3	16.7	17.8	18.0	10	20
	Surgeons in PPs	11.3	7.5	15.1	12.0	2	17
Suturing							
	Orthopaedics	15.1	13.7	16.4	15.0	10	19
	GI/general	18.1	17.4	18.8	19.0	14	20
	Vascular	17.1	15.7	18.5	18.0	12	19
	Urology	17.7	16.5	18.9	18.0	14	19
	Plastic	14.4	11.4	17.4	14.0	8	18
	Cardiothoracic	17.3	16.0	18.6	17.0	14	20
	All volunteers	16.8	16.3	17.4	17.0	8	20
	Surgeons in PPs	12.2	8.5	15.8	12.0	3	18
Dissection							
	Orthopaedics	17.1	16.0	18.3	18.0	14	20
	GI/general	18.8	18.1	19.6	20.0	14	20
	Vascular	19.8	19.4	20.2	20.0	18	20
	Urology	19.2	18.2	20.2	20.0	16	20
	Plastic	16.9	13.9	19.9	18.0	10	20
	Cardiothoracic	17.9	16.3	19.5	19.0	13	20
	All volunteers	18.4	17.9	18.8	19.0	10	20
	Surgeons in PPs	13.7	11.7	15.7	13.0	10	18

Table 2 (continued)

Simulation	Specialty	Mean	95% CI		Median	Range	
			Lower	Upper		Min	Max
Ligation							
	Orthopaedics	16.1	15.0	17.3	16.0	13	20
	GI/general	18.7	17.9	19.6	19.0	12	20
	Vascular	19.0	18.3	19.7	19.0	17	20
	Urology	19.6	19.1	20.1	20.0	18	20
	Plastic	16.4	15.5	17.3	16.0	15	18
	Cardiothoracic	18.5	17.5	19.6	19.0	16	20
	All volunteers	18.1	17.7	18.6	19.0	12	20
	Surgeons in PPs	15.4	14.4	16.5	15.0	14	18
Hand–eye							
	Orthopaedics	15.8	14.1	17.5	17.0	7	20
	GI/general	17.8	16.8	18.8	18.0	10	20
	Vascular	17.2	14.8	19.6	19.0	9	20
	Urology	17.8	16.9	18.7	18.0	15	20
	Plastic	15.7	14.6	16.9	15.0	15	18
	Cardiothoracic	17.0	14.6	19.4	18.0	7	20
	All volunteers	17.0	16.4	17.6	18.0	7	20
	Surgeons in PPs	13.6	9.2	17.9	14.0	5	19
All simulations							
	Orthopaedics	112.9	107.9	117.8	115.0	98	128
	GI/general	120.0	115.8	124.1	121.0	96	133
	Vascular	122.0	116.8	127.2	119.0	112	134
	Urology	124.9	120.2	129.6	125.0	111	134
	Plastic	113.7	107.7	119.7	116.0	101	119
	Cardiothoracic	123.3	119.1	127.5	122.5	112	134
	All volunteers	119.4	117.3	121.4	121.0	96	134
	Surgeons in PPs	93.8	80.8	106.9	98.0	70	119

The scores of surgeons in the PPs were significantly lower than the volunteer group for each simulation and overall ($P = 0.0001$) except for hand–eye co-ordination ($P = 0.004$), scrubbing and patient preparation (not significant).

Table 3 Summary of ANOVA

Effect	Sum of squares	df	Mean square	F	*P*-value	*Partial η^2
Simulation hypothesis	280.886	6	46.81	3.85	0.005	0.414
Error	397.302	32.666	12.16			
Specialty hypothesis	198.891	5	39.78	2.52	0.046	0.253
Error	588.220	37.324	15.76			
Persons within specialty hypothesis	596.731	68	8.77	1.52	0.008	0.202
Error	2362.502	408	5.79			
Specialty x Simulation hypothesis	383.248	30	12.77	2.21	0.000	0.140
Error	2362.502	408	5.79			

The dependent variable was the score on simulation. Because the surgeons were nested within a specialty group, each effect was calculated with reference to its own error term. *Partial η^2 is an independent measure of effect size.

recruiting cosultant surgeons to give a day's worth of time to this type of work cannot be overstated. No one test can be completely discriminatory and this study is part of a larger programme to establish the validity and reliability of such tests, including more advanced specialty-specific simulations. A time was allocated for each simulation for the volunteers because of the constraints of the OSCE format but the surgeons in the PPs were allowed unlimited time. Almost all volunteers completed each simulation within the allocated time but some of the surgeons in the PPs did not. The time taken to complete a task has been shown to correlate with performance.[10] The time to completion was not measured in the PPs because of the increased anxiety that this might have caused, but this may be an additional factor to consider in the future.

We believe that this is the first study to examine the differences in 'generic' skills between different surgical specialties. The significant inter-specialty ANOVA together with the comparable scores of most specialties suggest that some skills may not be generic for all surgeons. Nevertheless, taken together, this 'basket' of skills represents a fair test and shows large differences between surgeons potentially identified as poor performers and a volunteer sample from 6 surgical specialties. This is not surprising as all surgical specialties in the UK share a common basic surgical training. Less need for certain skills might explain the lower scores achieved by some surgeons in particular simulations. The different techniques used by 'conventional" and 'endoscopic' surgeons may explain the low scores of the three surgeons who did not undertake any endoscopic work. Certain specialties or individuals may not require some skills, or the skills acquired during earlier training may degrade through lack of use or emphasis.

Conclusions

This study confirms that relatively simple 'generic' simulations can be used to assess the technical skills of consultant surgeons. The lower scores of the surgeons in the PPs suggests that such tests possess criterion validity and that the Phase 1 assessments help to identify those doctors whose poor performance may be due to a lack of technical competence.

Acknowledgements

The authors thank Dr Tessa Dunseath and Sister Susan Cowley at the Clinical Skills Centre, Northern General Hospital and all the consultant volunteers for their help with the assessment exercises and Professor Richard Reznick for his assistance with the development of the task-specific checklists. The General Medical Council funded development of the simulations plus the travel expenses of the assessors and volunteers.

References

1. Southgate L, Cox J, David T, Hatch D, Howes A, Johnson N *et al*. The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's Performance Procedures. *Med Educ* 2001; **35 (Suppl)**: 2–8.
2. General Medical Council. *Maintaining Good Medical Practice*. London: GMC, 1998.
3. Rethans J, Sturmans F, Drop R, Van der Vleuten C, Hobus P. Does competence predict performance? Comparison between the examination setting and actual practice. *BMJ* 1991; **303**: 1377–80.
4. Thomas WEG. Core skills, courses and competency. *Ann R Coll Surg Engl* 2000; **82 (Suppl)**: 18–20.
5. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg* 1994; **167**: 423–7.
6. Beard JD, Jolly BC, Newble DI, Thomas WEG, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. *BMJ* 2005; Submitted.
7. Crick JE, Brennan RL. *Manual for GENOVA: a Generalized Analysis Of Variance System*. Iowa City: ACT Bulletin No. 43, 1983.
8. Paisley AM, Baldwin PJ, Paterson-Brown S. Validity of surgical simulation for the assessment of operative skill. *Br J Surg* 2001; **88**: 1525–32.
9. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchinson C. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997; **84**: 273–8.
10. Szalay D, MacRae H, Regehr G, Reznick R. Using operative outcome to assess technical skill. *Am J Surg* 2000; **180**: 234–7.