Multiple Objective Measures of Skill (MOMS)

A New Approach to the Assessment of Technical Ability in Surgical Trainees

Sean Mackay, MBBS, FRACS, Vivek Datta, MD, BSc FRCS, Avril Chang, MBBS, FRACS, Jyoti Shah, MD, BSc MRCS, Roger Kneebone, PhD, FRCS, FRCSEd, MRCGP, and Ara Darzi, MD, FRCS, FRCSI, FACS

Objective: The assessment of surgical technical skills has become an important topic in recent years. This study presents the validation of a 6-task skills examination for junior surgical trainees (at the level of the Membership of the Royal College of Surgeons).

Summary Background Data: Six tasks were evaluated in a project that also examined the feasibility of this method of assessment. The tasks were knowledge of sutures and instruments; knowledge of surgical devices; knot formation; skin-pad suturing, closure of an enterotomy; excision of a skin lesion; and laparoscopic manipulation. Comparisons were made between a group of junior trainees (n = 13), and a group of seniors (n = 8).

Results: Each of the 6 tasks was able to be used to discriminate between the 2 groups. In all, there were 19 primary analyses across the 6 tasks, and 17 of these showed significant differences between the groups (P values ranging from 0.037 to < 0.001). There was generally a strong correlation between the analyses, and when a mean rank was calculated, the difference between groups was significant ($P = 0.005$ on Mann–Whitney U test; mean ranks 13.9 and 6.3 [of 21], for juniors and seniors respectively). Reliability of the 6-task assessment was very good at 0.70 (Cronbach's Alpha).

Conclusions: A skills examination is a feasible and effective method of assessing the technical ability of basic surgical trainees.

(*Ann Surg* 2003;238: 291–300)

Technical competence in surgery has come under increased scrutiny in recent years. In the United Kingdom, most attention has focused on a few high profile instances involving consultant (attending) surgeons, in which it has been suggested that poor clinical outcomes were the result of inadequate technical skill. However, there has also been significant concern about standards for trainees. Because the changes brought about by the reduced hours of work (European directive)[1,2] for basic trainees are added to those already imposed on advanced training,[3] there have been suggestions that standards may suffer as training becomes shorter and less intense.[4] In the United States, the Accreditation Council for Graduate Medical Education (www.acgme.org) has addressed this broad issue in terms of both residents' hours of work and the definition and assessment of competencies.

A number of different techniques for the objective assessment of surgical skill have been proposed,[5] but, as yet, implementation outside the laboratory setting has been difficult. This article describes a project that aims to validate a multitask skills examination, aimed at basic surgical trainees.[6,7] The basic premise is that the process will be more robust if candidates are assessed on multiple parameters using a variety of measures.

Trainees at this level were chosen as the focus of this project for several reasons. Most importantly, the Membership of the Royal College of Surgeons represents a transition in training, from undifferentiated Basic Trainee, to Advanced Trainee on a recognized training scheme. Hence, it is a natural breakpoint at which to envisage a generic examination of technical competence. Beyond this, it is our opinion that the objective measures currently available are better suited to

From the Imperial College of Science, Technology and Medicine, Academic Surgical Unit, St. Mary's Hospital, Paddington, London, United Kingdom.

Supported by the Royal Australian College of Surgeons, via the Lumley Fellowship, a reciprocal RACS/RCS exchange fellowship (S.M.), and the BUPA Foundation (V.D.). A.C. is a National Health Service (NHS) consultant at Central Middlesex Hospital and an Honorary Senior Lecturer within Imperial College of Science, Technology and Medicine (ICSTM). J.S. is a research fellow within ICSTM and is funded by a grant from the NHS. R.K. is a part-time research fellow within ICSTM and is funded by Imperial College. A.D. is Professor of Surgery within ICSTM.

Reprints: S. Mackay, Imperial College of Science, Technology and Medicine, Academic Surgical Unit, 10th floor QEQM Building, St Mary's Hospital, South Wharf Road, Paddington, W2 1NY, London, United Kingdom. E-mail: s.mackay@ic.ac.uk.

Copyright © 2003 by Lippincott Williams & Wilkins

0003-4932/03/23802-0291
DOI: 10.1097/01.sla.0000080829.29028.c4

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

relatively junior rather than senior trainees. The UK and North American systems are not directly comparable; however, it is reasonable to equate basic trainees with residents in PGY 1 and advanced trainees with residents in PGY 3–5.

Ultimately, it may be possible to use such an assessment to determine which trainees have achieved a necessary level of performance. We suggest that such a process should be based around an assessment of competence rather than a competitive ranking system.[8]

The Multiple Objective Measures of Skill (MOMS) examination was based around the format of an Objective Structured Clinical Examination (OSCE),[9] as used in most medical schools. The aim was to develop 6 valid and complementary tasks, each taking 15 minutes, for a total examination time of 90 minutes. The tasks were defined in full before the commencement of data collection; hence, this article describes a validation rather than a pilot study. The tasks were based on previous work in the authors' department, bearing in mind the skills taught on the Basic Surgical Skills Course. This skills course is a compulsory part of basic surgical training and is jointly administered by the 4 Royal Colleges of Surgeons (www.rcseng.ac.uk; Royal College of Surgeons of England). The primary comparison planned for each task was between the junior group (basic trainees) as a group and their seniors (advanced trainees and attending surgeons), as that comparison respects the break-point in training mentioned above.

Subjects and Recruitment

Twenty-one subjects were recruited in all, composed of 13 juniors and 8 seniors. Eleven subjects were assessed individually in the skills laboratory, whereas the remaining 10 were assessed in 2 MOMS examinations held in the department (5 + 5 subjects). Each examination was designed around simultaneous assessment of 6 subjects, but on each occasion, 1 subject had to withdraw because of clinical commitments.

Statistical Methods

Previous experience in this area has demonstrated that the data are usually nonparametric. This is because there is typically a strong "floor" or "ceiling" effect, whereby there is a limit to how well a subject can perform using the measures available. This creates a skew in the data, which varies from task to task. Sometimes, it is be possible to normalize the data by log transformation; however, this technique cannot be applied universally.

Hence, it was determined that nonparametric analyses would be used throughout. The primary comparison is always between 2 groups (junior and senior) and the Mann–Whitney *U* test (MWU) is used in these comparisons. Where boxplots are presented to demonstrate the data graphically, the following apply: the heavy line is the median, the box represents the interquartile range (ie, 25th to 75th), and the "whiskers" represent the range of the data.

For the objective structured assessment of technical skill (OSATS) tasks, between-observers reliability was assessed by using Cronbach's coefficient alpha (α). Comparisons between the results of the 6 tasks were made by using Spearman's nonparametric correlation, and intertask reliability was assessed by using Cronbach's α. Cronbach's α quantifies the proportion of true score (rather than random error) in a summed scale. All statistics were calculated by using the Statistical Package for Social Sciences (SPSS, Chicago, IL) on a PC.

METHODS

Subjects were presented with a standardized instruction sheet before assessment and had this document available throughout the process for further reference. Each task was directly supervised, and one of the authors acted as supervisor for all data collection. All surgical procedures were performed on synthetic tissues (Limbs and Things Ltd, Bristol, UK).

Knowledge of Instruments, Sutures, and Surgical Devices (Task 1)

This task aimed to test familiarity with common surgical instruments and sutures, and with a laparoscopic insufflator and an electrosurgical device (diathermy). The component that dealt with sutures and instruments involved presenting the subjects with 8 color photographs (297 × 210 mm), each depicting (life-size) 3 or 4 instruments or sutures, each marked with a letter. The subjects were asked to answer a question on a printed examination paper as to which of the options on each photograph would be most appropriate under given circumstances. Four of the 8 questions dealt with sutures, and the other 4 with instruments. This component was marked out of 8 points.

The component that dealt with the surgical devices involved presenting the subjects with each of a laparoscopic insufflator and a diathermy (Electronic Endoflator 264300 20 and Autocon 350, both produced by Storz, Tuttlingen, Germany). The various connections and settings were marked with letters, and the subjects were asked to answer a series of questions relating to the safe use of the devices. This component was marked out of 17 points (8 for the insufflator and 9 for the diathermy). Hence, the total for all components of task 1 was 25 points. One of the subjects had worked in the skills laboratory on an unrelated project and potentially had knowledge of the marking schedule. He was excluded from this component of the study, which therefore involved 20 subjects rather than 21.

Knot formation (Task 2)

Subjects were asked to tie 4 surgical knots using a familiar proprietary jig. The 4 knots comprised 1 each of 10

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

single throws at the surface using cord, 10 throws at surface using 45-cm 2/O Polysorb ties (braided copolymer of glycolic and lactic acid, US Surgical, Norwalk, CT), 4 throws at depth using cord, and 4 throws at depth using 2/O Polysorb.

The Imperial College Surgical Assessment Device (ICSAD) was used to determine the number of movements and time required for each of these tasks. This device, previously described,[10–13] uses electromagnetic motion tracking and purpose-written software algorithms to compute the time and movement data. The knots were cut from the jig on each occasion and examined by the supervisors as failure to comply with the instructions could have confounded the measurement.

Skin-Pad Suturing (Task 3)

This task was likewise assessed by using the ICSAD. Although envisaged as a single exercise in the final MOMS format, 2 components were assessed during the validation process. Each involved placing 5 interrupted sutures in a synthetic skin-pad, the difference being that the 1 involved simple sutures and the other involved vertical mattress sutures. All knots were instrument-tied and involved 4 single throws of 3/O Polysorb.

Each task involved a single 4-cm incision with suture entry and exit points marked 1 cm back from the incision on both sides. The suture points were 1cm apart, giving 5 sutures over the 4-cm wound. It is appreciated that skin suture bites of 1 cm each side are unrealistically large, but this specification served to standardize technique—with the needle provided (PC12, 19 mm, 3/8 circle) it was not possible for a subject to incorporate both sides in a single bite (which would have confounded the motion analysis results). As for the knot tasks, all skin pads were retained for later inspection, to check for accuracy of sutures and of wound closure.

Enterotomy Closure (Task 4)

This task involved the closure of a 2-cm transverse enterotomy in synthetic small bowel. The standardized technique required the placement of stay sutures at each end, and the use of interrupted sutures, which were hand-tied, using 4 throws of 3/O Surgidac (braided polyester, US Surgical) on a V20 (26 mm, 1/2 circle, taper-point) needle. The synthetic bowel was positioned in a standard jig.

The performance of this task was videotaped by using digital video tape for later analysis using an OSATS technique.[14,15] In line with the published literature, and prior experience, both checklist and global scoring sheets were used for the assessment. The checklist involved 15 separate items, all of which required a yes/no answer for a maximum score of 15. The global score was composed of 8 parameters, which were scored from 1 (very poor) to 5, giving a maximum possible score of 40. All video data were assessed by 3 trained observers, all surgeons from within the department. The subjects were identified only by a number, and their faces were not seen on any of the footage. The soundtrack was turned off during the scoring process, to maximize anonymity. The final score for each subject was generated by summing the scores of the 3 observers, hence the maxima possible were 120 for the global assessment and 45 for the checklist.

Excision of a Skin Lesion (Task 5)

Video-based OSATS was again used in the assessment of this task, and the experimental setup and scoring process were as for the enterotomy task. The checklist again involved 15 items and the global scoring sheet was the same as used in the enterotomy task. The assessment was carried out by the same 3 observers. Maximum scores were again 120 (global score) and 45 (checklist).

The task itself involved the excision of a sebaceous cyst from synthetic skin. The standardized technique required the lesion to be excised as an ellipse and the wound closed with interrupted sutures of 3/O Polysorb, on a V20 needle, with all knots instrument-tied and comprising 4 throws. The skin pad was held in a standard jig.

Laparoscopic Task (Task 6)

The Minimally Invasive Surgical Trainer—Virtual Reality (MIST VR, Mentice, Gothenburg, Sweden[16]) was used for this component. This device has 6 available tasks, and 2 of these were chosen—the "acquire place" and "traversal." The acquire place task is a single-handed task that involves grasping an object in virtual space and then placing it in a defined position (wire-frame cage). The traversal task is a two-handed task that involves walking 2 graspers along a cylinder from one end to the other.

The program includes HTML-based files that serve to introduce, explain, and demonstrate each task, and these were used as a standardized introduction, along with the instruction sheets mentioned previously. The experimental set-up provided each candidate with the introduction/demonstration, followed by 3 consecutive trials each of which involved both a left and a right-handed repetition. The subjects were taken through the introduction and then assessment on Acquire Place, and then likewise for the Traversal component. The MIST VR has been validated as a test of laparoscopic skill.[17,18] In this setting, it was envisaged as a test of dexterity in the laparoscopic environment.

RESULTS

Knowledge of Instruments, Sutures, and Surgical Devices (Task 1)

There was a significant difference in median overall score out of 25 (Fig. 1). The junior group had a median score of 17, and the senior group a median score of 22 ($P = 0.024$). When the 2 components of this task (sutures and instruments, surgical devices) were examined separately, there were sig-

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

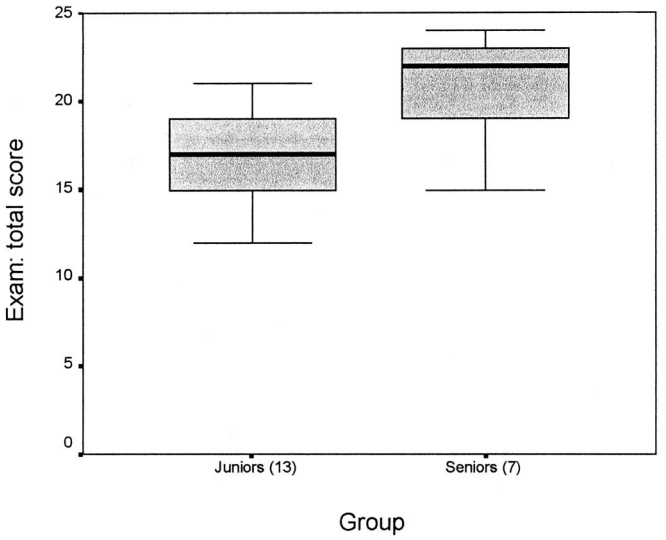

FIGURE 1. Total score for task 1 by group. Maximum possible score was 25. $P = 0.024$ (MWU).

nificant differences between the groups for each analysis (Table 1).

Knot Formation (Task 2)

There were significant differences between the groups for movement and time on 7 of the 8 measurements (Table 2). The greatest differences were observed for the knots formed at depth, and the nonsignificant difference was observed for the time component of the least difficult knot, the use of cord at the surface. The data for the best discriminator, 2/O Polysorb at depth, are presented graphically in Figure 2, which demonstrates a marked "floor" effect for the seniors' data.

Skin-Pad Suturing (Task 3)

There were significant differences between the 2 groups on all 4 parameters. Graphical analysis of the movement data for the 2 tasks (Fig. 3) demonstrates that the more difficult task, the mattress suture, has a broader spread of performance, and a higher median, for both groups (Table 3).

Enterotomy Closure (Task 4)

The global assessment was highly discriminatory ($P < 0.001$), but the checklist did not show a significant effect (Table 4). Figure 4 demonstrates that both groups had reached the ceiling of performance for the checklist assessment, which was therefore nondiscriminatory. Reliability between observers (Cronbach's α) was 0.88 for the global assessment and 0.82 for the checklist.

Excision of a Skin Lesion (Task 5)

Both components showed significant differences between the groups (Table 4), with the seniors performing better ($P = 0.008$ for global score; $P = 0.037$ for checklist). As for Task 5, the global score has a maximum of 120 and the checklist a maximum of 45. Reliability between observers (Cronbach's α) was 0.87 for the global assessment and 0.88 for the checklist.

Laparoscopic Task (Task 6)

Data for 2 subjects, 1 from each group, were corrupted by the data collection process, and these subjects have been excluded from further analysis on this component. The junior group show progressively lower scores for each of the 3 trials, and this learning effect was significant ($P = 0.003$, Wilcoxon Signed Rank Test). The observed differences between the junior and senior groups were analyzed on a trial by trial basis and each was significant ($P = 0.002$, 0.005, and 0.028 for trials 1, 2, and 3 respectively). When an average score was calculated for the 3 trials, the observed difference between groups (juniors' mean rank 8, seniors' 12 [of 19] was not significant ($P = 0.21$ MWU).

There was no difference between the 2 groups in terms of performance on the traversal task, and nor was there a learning effect for either group. This finding was repeated on subset analyses of the 3 components of the total score – time taken, economy of movement (the ratio of actual to ideal path-length), and error score.

TABLE 1. Overall and Subset Analyses of Task 1: Sutures & Instruments, Insufflator, and Diathermy

	Overall (Mean Rank)	Sutures and Instruments (Mean Rank)*	Insufflator and Diathermy (Mean Rank)*
Juniors (13 subjects)	17 (8.3)	6 (8.4)	10 (8.0)
Seniors (7 subjects)	22 (14.5)	7 (14.3)	14 (15.0)
P (MWU)	0.024	0.037	0.011

*Summary statistics are medians.
Mean ranks (of 20) are given in parentheses.

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

TABLE 2. Results for Knot Formation

	Cord		2/0 Polysorb	
	Movements (Mean Rank)*	**Time (Mean Rank)***	**Movements (Mean Rank)***	**Time in seconds (Mean Rank)***
Knot formation at the surface				
Juniors	69 (13.2)	30 (12.5)	62 (13.7)	32 (13.6)
Seniors	51 (7.4)	23.5 (8.6)	48.5 (6.6)	19.5 (6.8)
P (MWU)	0.037	0.19	0.008	0.013
Knot formation at depth				
Juniors	52 (14.1)	25 (13.7)	41 (14.3)	21 (14)
Seniors	35 (5.9)	16 (6.7)	26.5 (5.6)	12 (6.1)
P (MWU)	0.002	0.010	0.001	0.003

*Summary statistics are medians. Mean ranks (out of 21) are given in parentheses.

Correlations

Table 5 presents nonparametric (Spearman) correlations for selected components of each task. The components were chosen on the basis that they offered the best discrimination between subjects, within each task. It is true that the surgical devices component of the examination was actually a better discriminator than the total score, but that finding was the result of a subset analysis.

Interestingly, the acquire place component of task 6, although showing significant differences between groups and a significant learning effect for the juniors, failed to show a significant correlation with any other component of the MOMS examination (ie, not only those in the correlation matrix). When this analysis was repeated as a comparison between the 2 groups, to explore the possibility that there may be no correlation because of poor performances by the junior group, there were again no significant correlations between the acquire place tasks and any other parameter.

Overall Reliability

Overall reliability for the 6 tasks was assessed by using Cronbach's coefficient α, and the calculated value for α was 0.70. When using this test, it is necessary to use just 1 measure per task, as the method is sensitive to tight correla-

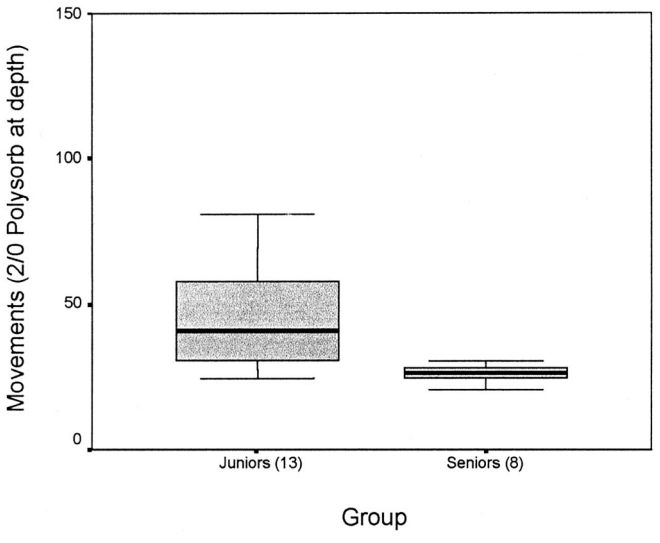

FIGURE 2. Movement data for the 2/O Polysorb-at-depth component of task 2 (knot formation). $P = 0.001$ (MWU). The position of a single outlier in the junior group (with a score of 200) is not shown in this graph.

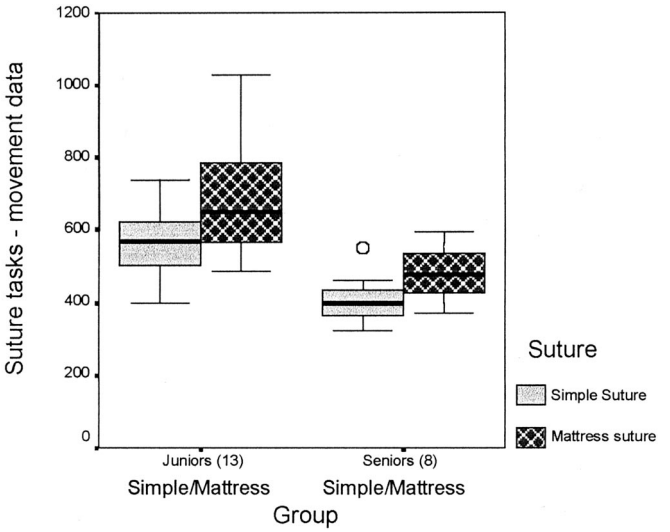

FIGURE 3. Movement data for the skin-pad suture task (task 3). The data for simple and mattress sutures are presented for each group. Differences between groups were significant ($P = 0.001$ for simple suture, and $P < 0.001$ for mattress suture; MWU). The single outlier for the simple suture task in the senior group is indicated.

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

TABLE 3. Data for the Simple and Mattress Suture Components of Task 3

Suture Tasks	Simple suture		Mattress Suture	
	Movements (mean rank)*	Time (secs) (mean rank)*	Movements (mean rank)*	Time in seconds (mean rank)*
Juniors	569 (14.4)	332 (13.4)	649 (14.5)	397 (13.9)
Seniors	399 (5.5)	257 (7.1)	477 (5.4)	300 (6.25)
P (MWU)	0.001	0.025	<0.001	0.005

*Summary statistics are medians. Mean ranks (of 21) are given in parentheses.

tions between 2 or more components of the same task (which are effectively repeated measures from the same subjects). For this reason, the data for the 6 tasks used in the correlation matrix were transformed thus: each score was expressed as a proportion of one, and, where there were 2 components to the task (such as time and movement for tasks 2 and 3) a mean was calculated between the 2 proportions.

Mean Rank (Fig. 6)

A mean rank across 18 analyses was obtained. The analyses were: total score on task 1, all 8 time and movement scores on task 2, all 4 time and movement scores on task 3, global and checklist scores for task 4, global and checklist scores for task 5, and the acquire place component of task 6. The data are presented graphically in Figure 6. The observed difference between groups was significant ($P = 0.005$ MWU). Where a data point was lacking, the subject was ranked last. This applied to 1 subject from the senior group in task one, and 2 subjects (one from each group) in task 2—this strategy supports (task one) or is neutral toward (task six) the null hypothesis. These missing data-points represent 3 points out of 378 used in determining the mean rank.

DISCUSSION

The aims of this project were two: to establish a viable format for a multiple assessment skills examination for basic surgical trainees (based around the skills taught on the Basic Surgical Skills Course) and to validate 6 complementary tasks.

In terms of the first objective, the 2 OSCE-format data collection sessions have demonstrated that these 6 tasks can be administered efficiently, using relatively modest amounts of equipment (1 MIST VR, 2 ICSAD, 2 digital video cameras, a single laparoscopic insufflator and a diathermy, and a set of simple instruments and sutures). These 2 OSCEs were conducted by 3 research staff, with a fourth acting as supervisor. All tasks were easily completed within the allowed time frame, even by junior trainees.

In terms of the second objective, the data obtained do suggest that the 6 tasks can discriminate between subjects of differing ability, although the laparoscopic task (Task 6) will require refinement. Comments will be offered on each task, and subsequently, on the integration of the tasks:

In task 1 (knowledge of instruments etc), the results obtained show that the questions posed do discriminate between the 2 groups. There is obviously potential to further refine this task, and that is the subject of ongoing work; in particular, efforts are being made to develop a wide panel of valid questions, to ensure that subsequent utility is not confounded by a body of "received wisdom". Of all the tasks, it seems that this would be the most susceptible to this form of confounding, as all the others involve a skilled performance to achieve a satisfactory score.

This study included knot formation (task 2) in 4 different formats. Significant differences between groups were seen in all formats, although only for movement data on the

TABLE 4. OSATS Data for Task 4 (Enterotomy Closure) and Task 5 (Excision of Skin lesion)

	Enterotomy Closure		Skin lesion	
	Global Assessment (Mean Rank)*	Checklist (Mean Rank)*	Global Assessment (Mean Rank)*	Checklist (Mean Rank)*
Juniors (13)	65 (7.5)	38 (9.6)	58 (8.2)	27 (8.8)
Seniors (8)	88 (16.8)	40 (13.3)	81 (15.5)	35 (14.6)
P (MWU)	<0.001	0.19	0.008	0.037

Maximum scores are 120 for global assessment and 45 for checklist assessment.

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

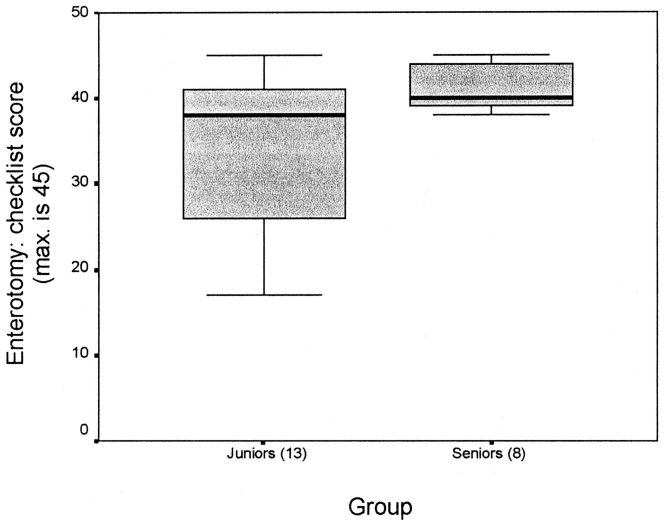

FIGURE 4. Checklist score results for task 4 (enterotomy closure). $P = 0.19$ (MWU).

"cord at surface" component. Intuitively, this is the easiest task, and it might therefore be expected that the juniors would, as a group, be closer to the "floor" level of performance. It is equally possible, however, that this "discrepant" result is simply a chance event, given that there are a total of 8 analyses under examination simultaneously.

Both skin-pad suturing tasks were highly effective in differentiating between the 2 groups. The senior group showed a strong "floor" effect (Fig. 3) for the easier of the two, the simple suture (which was done first on each occasion), but a broader spread of data for the mattress suture. This suggests that this task could serve to discriminate between individuals in the senior group, although the numbers studied here do not allow meaningful subset analysis

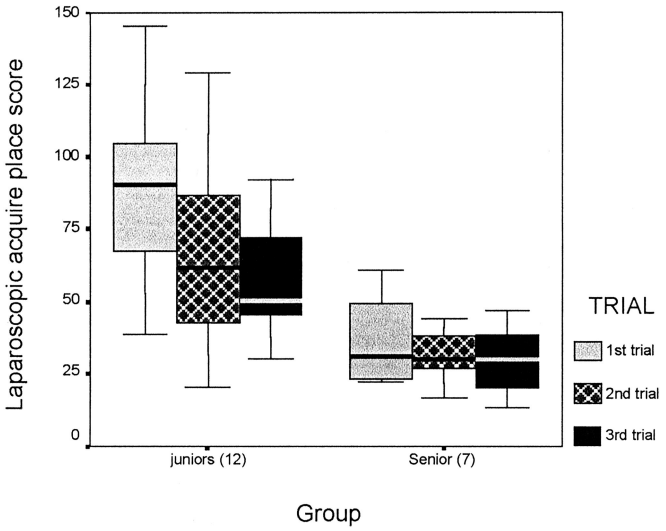

FIGURE 5. Trial-by-trial results for the MIST VR laparoscopic acquire place component of task 6.

The enterotomy closure (task 4) showed excellent reliability between the 3 observers for both global and checklist scoring. The global score was an effective discriminator whereas the checklist was not. Similar observations have been reported previously by others.[15,19] Examination of the graph of checklist score show that the seniors (median score 40 out of 45) were tightly clustered at the ceiling (Fig. 4), and that a significant proportion of the SHOs (median score 38) had also reached this level. Hence the measurement loses the power to differentiate. It is possible that "error detection" will improve the capacity to assess complex tasks, although this project did not attempt to address this question.

Both global and checklist scoring were significant discriminators on the excision of skin lesion (task 5), although the global was more effective (as assessed by relative position of median scores, the mean rank for each group, and the level of significance). On the global assessment, 7 of the 8 seniors (88%) bettered the 75th percentile for the juniors, whereas on the checklist, the interquartile ranges overlapped. There was not a strong ceiling effect for either group. Reliability between observers was again excellent for both global and checklist scoring.

The laparoscopic task (number 6) was the least successful of the 6 tasks. The acquire place task did differentiate between groups on a trial by trial basis, but not when the result of the 3 trials were averaged—this is because there was poor consistency to the ranking of individual subjects within the groups. Hence it is difficult to generate a valid summary statistic that can be used to generalize on the performance overall—the mean of 3 trials was used for the correlation calculations and there was no correlation with the rest of the MOMS examination. It is possible that this task really measures something different than the others. Likewise, it is also possible that the way the test was administered mitigated against maximum effectiveness.

This is the likely explanation for the complete failure of the traversal task, which has also been validated previously.[17,18] Intuitively, this seems to be the more difficult of the 2 used. It is possible that the subjects did not have enough time to acquaint themselves with the necessary performance prior to starting, or likewise, that they were all on a steep learning curve and the assessment was premature. It was a deliberate aspect of the MOMS design that the tasks involve minimal opportunity for practice, but perhaps that fact compromised this component.

Table 5 presents a correlation matrix for the components of each task that were seen to discriminate best between the 2 experimental groups. Apart form Task 6, there was good correlation between the MOMS tasks. Leaving aside task 6, Table 5 presents 21 different correlation analyses, of which 2 (both time and movement for mattress suture in task 3,

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

TABLE 5. Correlation Matrix

Correlation *P* value	1: exam total score	2: Polysorb at depth movement	2: Polysorb at depth time	3: Mattress suture movement	3: Mattress suture time	4: closure enterotomys global score	5: Skin lesion: global score	6: MIST acquire place
1: exam total score	NA	**−0.48** 0.033	**−0.50** 0.024	**−0.31** 0.18	**−0.36** 0.12	**0.52** 0.019	**0.66** 0.002	**−0.11** 0.64
2: 2/0 Polysorb at depth: movement	**−0.48** 0.033	NA	**0.94** <0.001	**0.85** <0.001	**0.85** <0.001	**−0.74** <0.001	**−0.61** 0.003	**0.20** 0.41
2: 2/0 Polysorb at depth: time	**−0.50** 0.024	**0.94** <0.001	NA	**0.83** <0.001	**0.86** <0.001	**−0.67** 0.001	**−0.59** 0.005	**0.13** 0.60
3: mattress suture: movement	**−0.31** 0.18	**0.85** <0.001	**0.83** <0.001	NA	**0.95** <0.001	**−0.59** 0.005	**−0.49** 0.024	**0.077** 0.75
3: mattress suture: time	**−0.36** 0.12	**0.85** <0.001	**0.86** <0.001	**0.95** <0.001	NA	**−0.61** 0.003	**−0.59** 0.005	**0.087** 0.72
4: anterotomy: global score	**0.52** 0.019	**−0.74** <0.001	**−0.67** 0.001	**−0.59** 0.005	**−0.61** 0.003	NA	**0.79** <0.001	**−0.43** 0.06
5: skin lesion: global score	**0.66** 0.002	**−0.61** 0.003	**−0.59** 0.005	**−0.49** 0.024	**−0.59** 0.005	**0.79** <0.001	NA	**−0.16** 0.50
6: MIST acquire place	**−0.11** 0.64	**0.20** 0.41	**0.13** 0.60	**0.077** 0.75	**0.087** 0.72	**−0.43** 0.06	**−0.16** 0.50	NA

Correlation coefficients are given in bold type and the *P* values are below them in standard type.

compared with task 1) did not show significance. There is no immediately obvious reason why this should be so, especially when all 3 measures correlated well with the others presented. It is possible that it is a chance event, given that multiple analyses are presented.

The overall reliability of the MOMS was very good at 0.70 (Cronbach's α). It was calculated that expanding the panel to 10 tasks (assuming equal utility) would raise α to 0.80, which is generally accepted as the benchmark for a high-stakes examination process.[20]

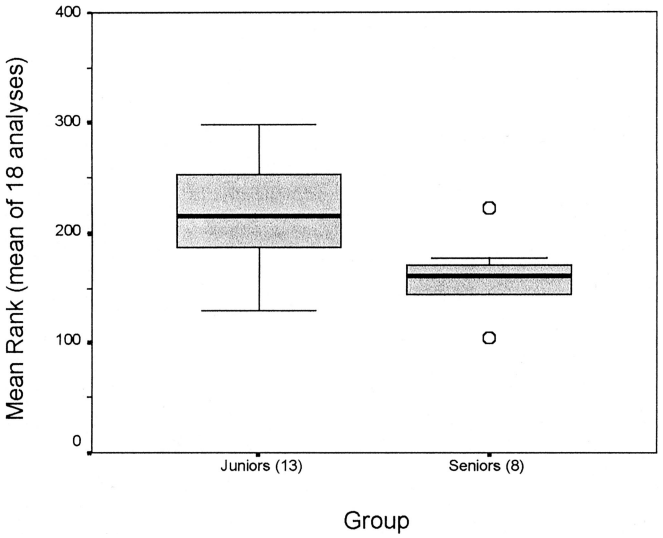

FIGURE 6. Mean rank across 18 analyses, between the 2 groups. $P = 0.005$ (MWU).

The mean rank takes the integration of the 6 tasks as 1 assessment a step further than the correlations presented above. In terms of methodology, the subset analyses of task 1 were not included, as they would have artificially increased the power; the time component for cord at surface in task 2 was included (despite being nonsignificant per se) as it was 1 of the planned components, and had the same pattern as the other 7 analyses of knot formation; and the traversal component of task 6 was excluded as it clearly needs reassessment before being used further in this setting. Figure 6 shows that 7 of 8 in the senior group (88%) performed better than 75% of the juniors—it might be possible to use such data to define a level of competent performance, once a larger database has been established.

The tasks used in this project were deliberately designed to be administered once only, to avoid the possibility that subjects would undergo a process of learning during the actual assessment. There was an improvement in performance on the acquire place component of task 6, which was statistically significant. It is possible that this improvement was due to increased familiarity with what was required, rather than a true learning effect, but it not possible to comment further on that possibility on the basis of these data.

The MOMS examination attempts to offer a multidimensional snapshot of a subject's ability, but specifically does not address issues of how quickly a subject may learn. Ability to learn is obviously important to the science of

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

surgical education,[21,22] but it was felt that a process of repeated trials would be too complicated and too time-consuming. Moreover, it is perhaps more appropriate to assess learning in a teaching setting, such as the Basic Surgical Skills Course, rather than an examination setting.

There were several examples of a strong floor or ceiling effect in the data gathered. The knot formation, skin-pad simple suturing, and enterotomy closure (checklist) tasks all showed this effect (Figs. 2, 3, and 4). This is not necessarily a disadvantage for the primary analyses (comparison of juniors and seniors), but does indicate that these measures would not be useful to discriminate between subjects in the better-performed group. The wider spread of data for seniors on the mattress suture compared with the simple suture (task 3) demonstrates how a more complex version of a given task will generally provide a broader spread of data, and hence be more useful in discriminating *within* the better performed group. It is equally true that these more difficult tasks may be of little or no value in assessing the junior group, as the subjects may simply find them too difficult to perform at all.

At this stage, 6 tasks have been evaluated. Task 1 is most susceptible to confounding by prior knowledge on the subjects' part, whereas the other tasks actually involve the performance of a fine motor task, and therefore mere knowledge of the content is unlikely to be a confounder. Significant prior practice would be a confounder if it were demonstrated that the ability gained through this practice did not generalize to other settings. At this stage there is a dearth of evidence in the medical/ surgical literature on this topic, and it is an area that warrants further study.

This project was first reported in abstract form in 2001[23] when preliminary analyses were presented. There is an extensive literature that deals with multitask skills assessments in which all tasks use the same methodology,[14,15,19,20,24–26] but there are very few publications that describe the use of multiple different indices of skill.

Cerilli and coworkers did publish (in 2001) a work that describes an OSCE that involved 3 technical skills stations and 5 clinical skills stations.[27] No reliability statistics are quoted. The authors concentrate on the correlation between year of training (PGY 1–5) and score. For technical tasks, the correlation was 0.679, and for clinical tasks it was 0.203, and this difference (between the pattern on technical and clinical skills stations) was statistically significant ($P < 0.05$). The authors do not attempt to generate a summed score and, as mentioned, there is no assessment of reliability.

Also in 2001, Paisley and coworkers[28] described an assessment process that involved 6 previously validated measures of skill, in basic surgical trainees in the UK (equivalent to PGY 1). The primary analysis was an assessment of construct validity made by comparing trainee performance on the 6 tasks, with an assessment of competence made by the trainees' supervising attending surgeons (using a previously validated assessment form). However of 12 measures, only 1 showed a statistically significant correlation, which was only −0.27. Task performance data were then correlated with the duration of basic training for each subject and 2 of 12 correlations were statistically significant, however both were weak correlations (−0.23, and −0.24).

The authors do not report a pilot phase, in which they demonstrate their facility with the tasks (which were developed by others) and therefore it is possible that there was some difficulty in the application of the measurement techniques. It is also true that the reference point for this attempt at construct validation was the assessment by the attending surgeons, which is itself a surrogate end point. Moreover, given that the subjects spanned a narrow range of experience, it is to be expected that they would span a relatively narrow band of ability. If the range of possible responses for both the reference standard and the experimental tasks is small, then the ability to discriminate between subjects (and hence to show correlations between measures) will be diminished.

Our department has prepared a detailed costing for the administration of the MOMS examination on a national basis in the UK. The costing was based around a single national testing center, with a projected throughput of 1000 candidates annually. Based upon this model, the set-up costs would be $75000 (not including the purchase of a venue, and/or any necessary refit), and the examination could then be offered to candidates for $200.

Hence, it is possible to measure performance on standardized tasks in an examination setting, using multiple objective measures. The challenge that springs from this is the interpretation of the data, and in particular, the assessment of what is a competent level of performance. The authors believe that, if an examination body were to introduce a process of this nature, it would be necessary to review the first 1 or 2 years of data to establish these definitions of competence.

The authors submit that the MOMS approach described here, in which several methods of assessment are used, has the potential to minimize any disadvantage to a given candidate that may arise by relying on one method of assessment across a multitask appraisal.

ACKNOWLEDGMENTS

The authors acknowledge the statistical advice given by Professor Glenn Regehr, PhD, of the Centre for Research in Education, Faculty of Medicine, University of Toronto, and support in this project and advice on the presentation of the manuscript given by Mr. R. C. G. Russell.

REFERENCES

1. NHS Management Executive. *Junior Doctors: the New Deal*. London: Department of Health; 1991.
2. Last GC, Curley P, Galloway JM, et al. Impact of the New Deal on vascular surgical training. *Ann R Coll Surg Engl*. 1996;78(6 suppl):263–266.

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

3. Department of Health. Hospital doctors: training for the future. The report of the working group on specialist medical training. London: Department of Health; 1993.
4. Skidmore FD. Junior surgeons are becoming deskilled as result of Calman proposals [letter]. *BMJ*. 1997;314:1281.
5. Reznick RK. Teaching and testing technical skills. *Am J Surg*. 1993; 165:358–361.
6. Examinations Board of the Royal College of Surgeons of England. *MRCS Regulations 1999*. 2 ed. London: RCS/Ashford Colour Press; 1999.
7. Training Board of the Royal College of Surgeons of England. *The Manual of Basic Surgical Training*. London: RCS/Sarcombe Press; 1998.
8. McManus IC. Examining the educated and the trained. *Lancet*. 1995; 345:1151–1153.
9. Cuschieri A, Gleeson FA, Harden RM, et al. A new approach to a final examination in surgery. Use of the objective structured clinical examination. *Ann R Coll Surg Engl*. 1979;61:400–405.
10. Taffinder N, Smith SG, Huber J, et al. The effect of a second-generation 3D endoscope on the laparoscopic precision of novices and experienced surgeons. *Surg Endosc*. 1999;13:1087–1092.
11. Datta V, Mackay S, Mandalia M, et al. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg*. 2001;193:479–485.
12. Datta V, Mandalia M, Mackay S, et al. Relationship between skill and outcome in the laboratory-based model. *Surgery*. 2002;131:318–323.
13. Datta VK, Mackay SD, Gillies D, et al. Motion Analysis in the Assessment of Surgical Skill. *Computer Methods Biomechanics Biomed Eng*. 2001;4:515–523.
14. Faulkner H, Regehr G, Martin J, et al. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med*. 1996;71:1363–1365.
15. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84: 273–278.
16. Sutton C, McCloy R, Middlebrook A, et al. MIST VR. A laparoscopic surgery procedures trainer and evaluator. *Stud Health Technol Inform*. 1997;39:598–607.
17. Chaudhry A, Sutton C, Wood J, et al. Learning rate for laparoscopic surgical skills on MIST VR, a virtual reality simulator: quality of human-computer interface. *Ann R Coll Surg Engl*. 1999;81:281–286.
18. Wilson MS, Middlebrook A, Sutton C, et al. MIST VR: a virtual reality trainer for laparoscopic surgery assesses performance. *Ann R Coll Surg Engl*. 1997;79:403–404.
19. Regehr G, MacRae H, Reznick RK, et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73:993–997.
20. Reznick R, Regehr G, MacRae H, et al. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173:226–230.
21. Gallagher AG, McClure N, McGuigan J, et al. Virtual reality training in laparoscopic surgery: a preliminary assessment of minimally invasive surgical trainer virtual reality (MIST VR). *Endoscopy*. 1999;31:310–313.
22. Macmillan AI, Cuschieri A. Assessment of innate ability and skills for endoscopic manipulations by the Advanced Dundee Endoscopic Psychomotor Tester: predictive and concurrent validity. *Am J Surg*. 1999; 177:274–277.
23. Mackay S, Datta V, Morgan P, et al. Competence day: development of a panel of objective assessment tasks for senior house officers at MRCS level. *Br J Surg*. 2001;88(suppl 1):46.
24. Regehr G, Freeman R, Hodges B, et al. Assessing the generalizability of OSCE measures across content domains. *Acad Med*. 1999;74:1320–1322.
25. Szalay D, MacRae H, Regehr G, et al. Using operative outcome to assess technical skill. *Am J Surg*. 2000;180:234–237.
26. MacRae H, Regehr G, Leadbetter W, et al. A comprehensive examination for senior surgical residents. *Am J Surg*. 2000;179:190–193.
27. Cerilli GJ, Merrick HW, Staren ED. Objective Structured Clinical Examination technical skill stations correlate more closely with postgraduate year level than do clinical skill stations. *Am Surg*. 2001;67: 323–327.
28. Paisley AM, Baldwin PJ, Paterson-Brown S. Validity of surgical simulation for the assessment of operative skill. *Br J Surg*. 2001;88:1525–1532.

© 2003 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.