Implementing workplace-based assessment across the medical specialties in the United Kingdom

James R Wilkinson,[1] James G M Crossley,[2] Andrew Wragg,[1] Peter Mills,[1] George Cowan[1] & Winnie Wade[1]

OBJECTIVES To evaluate the reliability and feasibility of assessing the performance of medical specialist registrars (SpRs) using three methods: the mini-clinical evaluation exercise (mini-CEX), directly observed procedural skills (DOPS) and multi-source feedback (MSF) to help inform annual decisions about the outcome of SpR training.

METHODS We conducted a feasibility study and generalisability analysis based on the application of these assessment methods and the resulting data. A total of 230 SpRs (from 17 specialties) in 58 UK hospitals took part from 2003 to 2004. Main outcome measures included: time taken for each assessment, and variance component analysis of mean scores and derivation of 95% confidence intervals for individual doctors' scores based on the standard error of measurement. Responses to direct questions on questionnaires were analysed, as were the themes emerging from open-comment responses.

RESULTS The methods can provide reliable scores with appropriate sampling. In our sample, all trainees who completed the number of assessments recommended by the Royal Colleges of Physicians had scores that were 95% certain to be better than unsatisfactory. The mean time taken to complete the mini-CEX (including feedback) was 25 minutes. The DOPS required the duration of the procedure being assessed plus an additional third of this time for feedback. The mean time required for each rater to complete his or her MSF form was 6 minutes.

CONCLUSIONS This is the first attempt to evaluate the use of comprehensive workplace assessment across the medical specialties in the UK. The methods are feasible to conduct and can make reliable distinctions between doctors' performances. With adaptation, they may be appropriate for assessing the workplace performance of other grades and specialties of doctor. This may be helpful in informing foundation assessment.

KEYWORDS evaluation studies [publication type]; multicentre study [publication type]; feasibility studies; clinical competence/*standards; medical staff, hospital/*standards; *specialties, medical; professional practice; ANOVA.

*Medical Education 2008: **42**: 364–373*
doi:10.1111/j.1365-2923.2008.03010.x

INTRODUCTION

In the UK all higher medical trainees (specialist registrars, SpRs) undergo an annual record of in-training assessment (RITA) to summarise the preceding year's assessments and to ensure they are competent to continue training or to certify as independent specialists. Prior to this study the three Royal Colleges of Physicians (RCP) had not approved any methods for assessment of SpRs. Previous assessments in use were applied on a local basis and varied in their quality, making the RITA process highly subjective and informal.[1]

Assessment of doctors' performance has become an important issue as a result of high profile cases of malpractice and the redesign of medical training. Assessing doctors in an honest and objective manner is a fundamental part of the General Medical Council's guidance document *Good Medical Practice*.[2] Assessing competence (what doctors do in controlled representations of practice) does not reliably predict performance (what doctors do in real life).[3]

In 2002 the RCP launched new curricula for all medical specialties. Following this there was a need to

[1]Royal College of Physicians, London, UK
[2]Academic Unit of Medical Education, University of Sheffield, Sheffield, UK

Correspondence: Dr James R Wilkinson, 65 Rogers Road, London SW17 0EB, UK. Tel: 00 44 7931 703058; Fax: 00 44 208 767 6546; E-mail: stigassessment@mac.com

Overview

What is already known on this subject

There is a need for feasible and reliable methods to assess doctors' performance in the UK.

There are no existing validated methods for the assessment of a variety of procedural skills on patients.

What this study adds

We describe a novel, feasible and reliable way of assessing procedural skills (DOPS).

The mini-clinical evaluation exercise (mini-CEX), DOPS and multi-source feedback are feasible to conduct on a national scale and can make reliable distinctions between doctors with appropriate sampling.

Suggestions for further research

Fully crossed studies looking at judge error are necessary for the mini-CEX and DOPS.

develop feasible and reliable methods of performance assessment. We aimed to assess three aspects of performance: the clinical encounter; practical procedural skills, and professional behaviours and performance.

The clinical encounter

For this we selected the mini-clinical evaluation exercise (mini-CEX), which was developed as a method of assessing clinical skills by direct observation.[4] An assessor observes a trainee consulting with a patient in a real clinical situation, such as an outpatient clinic, and scores the trainee on a form using pre-defined criteria. It is designed to assess a variety of skills, such as history taking, clinical examination and management. The trainee receives instant feedback after each assessment. This has been shown to be a reliable way of assessing the clinical skills of trainees in the USA, where it is now in widespread use.[4]

Practical procedural skills

The RCP could find no suitable pre-existing assessment tools for this purpose so developed a novel method called 'directly observed procedural skills' (DOPS). An assessor observes a trainee performing a practical procedure on a patient, from start to finish, and scores the trainee against pre-defined criteria. The DOPS assessment is similar in principle to the mini-CEX, with the exception that the whole procedure is observed in DOPS, whereas in the mini-CEX it is not necessary to observe the entire patient encounter.

Professional behaviours and performance

Peer assessment has been shown to be a reliable way of assessing doctors.[5] Many investigators now combine the ratings of several different staff groups into a single measure, called multi-source feedback (MSF) or 360-degree assessment. Previous work has used between 10 and 15 people to assess doctors, depending on the vocational mix of the assessors.[5,6] Prior to this study there was no formal assessment of medical SpRs' behaviours, yet this is one of the main reasons for them failing the RITA.[7]

There are no published data for the reliability, validity and feasibility of the mini-CEX or DOPS for UK trainees . There are some data for the feasibility and reliability of MSF assessment of paediatric junior doctors, but none for medical SpRs.[8] Similar assessment tools have been introduced in Canada (by the CanMEDS Project) and the USA (by the American Board of Internal Medicine).[9,10]

The primary aim of this study was to assess the feasibility and reliability of these methods for assessing UK medical SpRs, when implemented with minimal resources and training. In addition, we examined confounding factors that might affect validity.

METHODS

The study was an RCP initiative. Full details of ethical approval, form design, standard setting, study set-up and examples of the forms are available in the online supplementary material.

Trainees were responsible for initiating assessments and returning all completed paperwork. Assessors were provided with written instructions on how to use the tools to assess trainees but received no formal training. The DOPS assessment was piloted for cardiac catheterisation, endoscopy, neurophysiology studies and renal biopsy. A generic DOPS form was used in parallel with a procedure-specific DOPS form for each assessment. Over a 4-month period all

trainees underwent MSF assessment and either mini-CEX or DOPS assessments; some underwent all three methods of assessment. We aimed to recruit 100 trainees for each method.

Each trainee was required to complete one MSF assessment, where they were asked to choose five raters from each of four groups, consisting of: allied health professionals (AHPs); clerical or secretarial staff; doctors, and nurses. Raters returned completed forms directly to the trainees' educational supervisors (ESs), who summarised all the forms by calculating the range and mean score for each item on the MSF form. The summary was used to provide trainees with anonymous feedback without showing them individual responses. Each trainee was asked to complete 6–8 mini-CEXs and/or 6–8 DOPS assessments. They were also asked to seek assessments by as wide a range of assessors as possible.

Finally, each trainee met with his or her ES to be given feedback on all assessments. Each party then completed a questionnaire for each method. These collected simple demographic data, asked specific closed questions about the method, and contained an area allowing open comments.

Questionnaire data was entered manually into EXCEL spreadsheets. Assessment forms were scanned into an EXCEL spreadsheet using the Teleform Optical Reading System. Each scan result was manually verified.

Data analysis

Questionnaire data

Results from direct questions about the methods are expressed as the crude percentage of responses. Free text comments were analysed by identifying themes raised by more than 10% of responders.

Feasibility data

The time taken to complete each assessment, recorded on every assessment form, is expressed as a mean and range. Further feasibility data were collected from the questionnaires.

Reliability data

This is an indication of how consistent or reproducible the observed differences between trainees are. Generalisability theory was used to model the reliability of scores with different numbers of assessors and encounters or procedures.[11] The mean score across the whole of each instrument was used for this analysis as the instruments are designed to cover every important aspect of performance in their domain. A variance component analysis (VARCOMP in SPSS Version 11 [SPSS Inc, Chicago, IL, USA], using the MINQUE procedure) quantified the factors influencing the scores (such as individual assessor variation). We used two regression models,[12] which are described in detail in the supplementary material.

Validity and confounding factors

Several confounding factors for each method have the potential to influence their validity. For example, judges using the mini-CEX may consistently rate difficult consultations differently from easy ones. The significance of several potentially confounding effects was tested using an independent-samples *t*-test for binary effects (such as inpatient versus outpatient) and a 1-way ANOVA for effects with ≥ 3 alternatives (e.g. year of training) in SPSS (Version 11).

RESULTS

Seventeen of 23 medical specialties participated, with participants from all UK regions. Six specialties declined (representing 12.0% of all trainees).

The numbers of trainees who agreed to participate were 247, 177 and 331 for mini-CEX, DOPS and MSF assessments, respectively. Completed assessments were returned by 128 (52%), 59 (33%) and 230 (69%) trainees for each of these methods, respectively. Trainees completed a mean of 5.14 mini-CEX assessments (range 1–10, median 5) and 4.83 DOPS assessments (range 1–14, median 4). A mean of 12.4 raters assessed each trainee (range 1–21, median 16).

Generic scores on the DOPS correlated well with procedure-specific scores (R = 0.84) and were used for the subsequent reliability analysis as they allow the data from different procedures to be combined.

Feasibility

The low completion rate highlights feasibility problems. Unstructured telephone interviews with local study co-ordinators revealed that lack of time was the main factor preventing completion.

Mean observation time for mini-CEX assessments was 18.5 minutes (range 1–90 minutes, median

15 minutes) and mean time spent providing feedback was 6.8 minutes (range 1–75 minutes, median 5 minutes). Mean observation time for DOPS varied according to the procedure assessed; on average feedback time took an additional 20–30% of the procedure observation time. Mean time taken to complete each MSF assessment form was 5.7 minutes (range 1–10 minutes, median 5 minutes).

Table 1 summarises responses to the closed questions on the questionnaire. It shows that the majority of trainees and ESs considered all three methods of assessment to be practical. (The mini-CEX was the least well rated in this respect.) The majority of ESs did not report any conflict between their roles as clinician and assessor when conducting mini-CEX or DOPS assessments. Table 2 summarises the free text comments. A significant proportion of trainees and ESs found the methods time-consuming (especially the mini-CEX). Some ESs found the process of collating MSF data from individual forms to provide a summary sheet added a considerable additional administrative workload.

Table 1 Questionnaire results showing educational supervisors' and trainees' views of assessment methods

	Method					
	Mini-CEX		**DOPS**		**MSF assessment**	
Direct question asked	**ESs' views (%)**	**SpRs' views (%)**	**ESs' views (%)**	**SpRs' views (%)**	**ESs' views (%)**	**SpRs' views (%)**
Do you think this assessment method is practical?						
Yes	78	69	93	86	81	87
No	21	30	7	8	16	12
Don't know	1	1	0	6	3	1
Do you think the responses are a fair assessment of an SpR's ability?						
Yes	82	84	91	86	83	81
No	15	10	3	3	14	14
Don't know	3	6	6	11	3	5
Do you think the process is helpful to an SpR's personal development?						
Yes	87	80	91	74	75	75
No	10	15	6	20	23	24
Don't know	3	5	3	6	2	1
Has the process provided any useful information about the trainee that you did not know?						
Yes	40	79	44	63	25	75
No	59	19	53	29	73	24
Don't know	1	2	3	8	2	1
Has this assessment method provided any additional information to your opinion as an ES about individual trainees?						
Yes	37	Not asked	47	Not asked	30	Not asked
No	62		44		67	
Don't know	1		9		3	
If you were the consultant responsible for the patient's care, were you able to act objectively as an assessor without any conflict of interests?						
Yes	84	Not asked	81	Not asked	Not asked	Not asked
No	14		3			
Don't know	2		16			

Mini-CEX = mini-clinical evaluation exercise; DOPS = directly observed procedural skills; MSF = multi-source feedback; ES = educational supervisor; SpR = specialist registrar

Reliability

Table 3 presents the variance component estimates based on the sophisticated regression models. On the basis of these effects an assessment strategy for medical SpRs was recommended as follows.

1 Mini-CEX: a trainee should be observed by at least eight different assessors observing at least two encounters each. This reduces the significant error caused by V_a and V_{t*a}. It simultaneously reduces V_c.
2 DOPS: a trainee should be observed by at least three different assessors observing at least two procedures each. This adequately controls the more modest and even error that results from the same effects.
3 MSF: a trainee should be rated by at least 12 different assessors. This adequately controls the error caused by V_a and V_{t*a}.

Table 4 presents the results of the simple regression analysis showing how true and error variances combine to produce 95% confidence intervals (CIs) for scores with varying sample sizes. The CI narrows as more observations, ratings and performances are sampled. With the recommended assessment strategies (above) the CI for the three methods reduces to 0.33 for the mini-CEX, 0.54 for DOPS, and 0.48 for MSF. Figure 1 presents these CIs in relation to the distribution of actual scores on the scale. It shows that the chosen assessment strategies allow each assessment method to score all the doctors in this sample as better than unsatisfactory (1–3) with 95% confidence. However, the CIs cross at least two quartiles of doctors. This means that doctors' rankings should be interpreted with caution.

Validity and confounding factors

Regarding face validity, a majority of participants felt that the assessment methods were fair (Table 1). With each method, more senior trainees received significantly higher scores (mini-CEX: $F = 9.5$, $P < 0.05$; DOPS: $F = 14.1$, $P < 0.05$; MSF: $F = 4.3$, $P < 0.05$). Furthermore, a significant number of participants commented on the formative value of the assessments by indicating that they 'provided useful basis for feedback and discussion' or 'improved training' (Table 2).

Table 2 Main themes identified from questionnaire free text feedback for each method

	Percentage of consultants who made comments	Percentage of SpRs who made comments
Mini-CEX		
Provided useful basis for feedback/discussion	10	22
Method said to be valid	15	0
Time-consuming and/or administrative workload	46	46
Created an artificial setting	10	20
DOPS		
Provided useful basis for feedback/discussion	17	19
Method said to be valid	19	0
Value of formalised assessment process	23	0
Improved training as a result	0	27
Time-consuming and/or administrative workload	23	19
MSF assessment		
Provided useful basis for feedback/discussion	13	0
Time-consuming and/or administrative workload	22	17
Concern about self-selection of raters	0	13
Use of method should be confined to poorly performing trainees	11	4

SpR = specialist registrar; mini-CEX = mini-clinical evaluation exercise; DOPS = directly observed procedural skills; MSF = multi-source feedback

Table 3 Variance component estimates

	Estimate	Degrees of freedom	Proportion of overall variance (%)
Mini-CEX			
Trainee-to-trainee variation (V_t)	0.13	127	15
Assessor-to-assessor variation (V_a)	0.19	87	21
The consistent preference of an assessor for a particular trainee (V_{t*a})	0.32	20	36
Encounter-to-encounter variation (nested) ($V_{residual}$)	0.24	422	28
DOPS			
Trainee-to-trainee variation (V_t)	0.40	58	36
Assessor-to-assessor variation (V_a)	0.36	50	33
The consistent preference of an assessor for a particular trainee (V_{t*a})	0.15	11	14
Procedure-to-procedure variation (nested) ($V_{residual}$)	0.19	165	17
MSF			
Trainee-to-trainee variation (V_t)	0.22	229	23
Assessor-to-assessor variation (V_a)	0.40	2345	42
The consistent preference of an assessor for a particular trainee ($V_{residual}$)	0.33	436	35

Mini-CEX = mini-clinical evaluation exercise; DOPS = directly observed procedural skills; MSF = multi-source feedback

For the mini-CEX, case complexity influenced scores, with cases of low, medium and high complexity receiving mean scores of 7.29, 7.60 and 7.70, respectively ($F = 7.9$, $P < 0.05$). In addition, the setting influenced the scores, with inpatient and outpatient encounters receiving mean scores of 7.74 and 7.44, respectively ($t = 4.3$, $P < 0.05$). Previous experience with the mini-CEX, tone of consultation ('good news' versus 'bad news'), familiarity with the patient, the focus of the mini-CEX ('data gathering', 'diagnosis', 'management', 'counselling' and 'mixed'), and observation time did not influence mini-CEX scores.

For the DOPS assessment, only the procedure type was examined as a confounding factor. This had no significant impact on scores, with cardiac, endoscopic, neurophysiological and renal procedures receiving mean scores of 7.34, 7.73, 7.36 and 7.33, respectively ($F = 2.6$, $P > 0.05$).

For MSF, the different genders gave different mean scores, with male and female raters giving mean scores of 7.78 and 7.97, respectively ($t = 4.6$, $P < 0.05$). The different professional groups also gave different mean ratings, with AHPs, consultants, nurses, clerical staff, junior trainees and peers giving mean scores of 7.63, 7.57, 8.00, 8.14, 8.14 and 7.86, respectively ($F = 22.2$, $P < 0.05$). The age of the rater did not affect the score given.

DISCUSSION

Main findings

Regarding feasibility, this report shows that it is probably possible to implement workplace assessment on a national scale across the majority of the medical specialties. However, there were low response rates (although this is typical in voluntary workplace assessment[13]). A significant minority of participants responded 'No' to the question 'Do you think this assessment method is practical?' (especially with regard to the mini-CEX). Open comments indicated that participants found the mini-CEX assessment and MSF result synthesis to be time-consuming. Even amongst those who did participate, only the minority managed to obtain the suggested six mini-CEX assessments, six DOPS assessments or 20 MSF ratings. Clearly, without adequate time and resources the feasibility of these methods would be significantly reduced.

Regarding reliability, the simple regression analysis produces 'true' and 'error' effect sizes almost identical

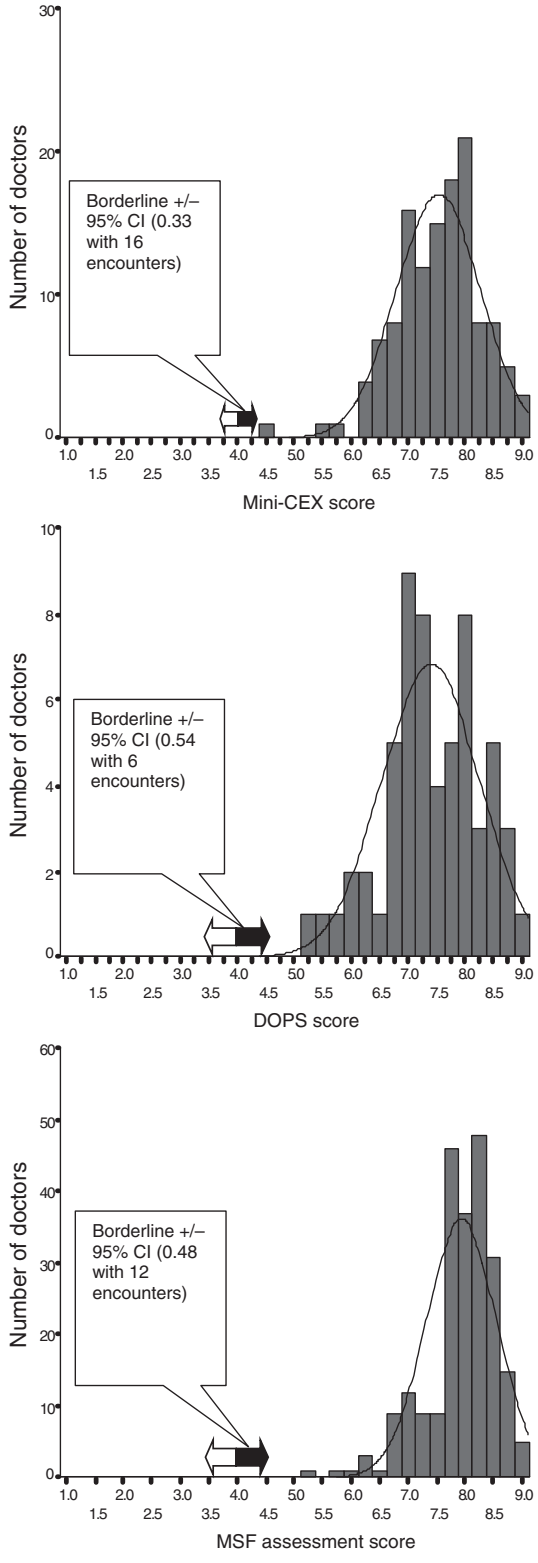

Figure 1 Histograms of scores, 95% confidence intervals (CIs) for a score of 4 (borderline) using the number of assessments recommended by the Royal Colleges of Physicians for each method. Mini-CEX = mini-clinical evaluation exercise; DOPS = directly observed procedural skills; MSF = multi-source feedback

to those found in previously published work for the mini-CEX[4] and for MSF.[8] This is the first published reliability evaluation of the DOPS assessment, and the data suggest that it compares favourably with the other two methods in that it requires fewer observations than the mini-CEX to achieve the same level of reliability. If trainees do not complete the number of assessments recommended by the RCP, the reliability of the methods is adversely affected. The findings of the more sophisticated regression analysis are discussed below.

Regarding validity, the majority of participants considered the methods fair. There were also many positive comments from trainees and ESs about the formative value of the assessments. The significant positive relationship between the seniority of trainees and their scores provides new evidence of the validity of the assessment methods. Importantly, DOPS scores did not appear to depend upon the procedure group in this population. However, the investigation of confounders showed some unwanted effects which have been observed before. For example, assessors' over-compensation for 'difficult' cases has been noted previously.[4] Further, newly observed confounders included the setting for the mini-CEX and the gender and professional designation of the rater for MSF.

Strengths and limitations

The voluntary mode of participation and the low response rates limit the generalisability of the findings. Volunteers are more likely to consider the methods fair and practical. It would also be reasonable to speculate that volunteers would obtain more favourable scores, so the results shown here cannot be considered as normative data. However, reliability is more difficult to achieve in uniformly high performers. Thus, it is likely that the reliability estimates are pessimistic and that smaller sample sizes would be sufficient to identify poor performers.

Controversies

The more sophisticated regression analysis has been used less often in published work.[14] However, now that procedures are available to estimate variance components on severely unbalanced data, this type of analysis has important advantages over simplistic analysis. For example, where an error effect (such as assessor) is heavily confounded with the subject of interest (trainee), the simplistic analysis will attribute a proportion of assessor variation to trainee-to-trainee variation. This will overestimate reliability and underestimate the sample size necessary to achieve a

Table 4 D study for different numbers of encounters or raters, showing generalisability coefficients, standard errors of the mean and 95% confidence intervals

Number of encounters or raters	Mini-CEX			DOPS			MSF		
	G	SEM	95% CI	G	SEM	95% CI	G	SEM	95% CI
1	0.45	0.68	1.34	0.56	0.67	1.32	0.23	0.85	1.67
2	0.62	0.48	0.95	0.72	0.47	0.93	0.38	0.60	1.18
3	0.71	0.39	0.77	0.80	0.39	0.76	0.48	0.49	0.97
4	0.77	0.34	0.67	0.84	0.34	0.66	0.55	0.43	0.84
5	0.80	0.31	0.60	0.87	0.30	0.59	0.61	0.38	0.75
6	0.83	0.28	0.55	0.89	0.27	0.54	0.65	0.35	0.68
7	0.85	0.26	0.51	0.90	0.25	0.50	0.68	0.32	0.63
8	0.87	0.24	0.47	0.91	0.24	0.47	0.71	0.30	0.59
9	0.88	0.23	0.45	0.92	0.22	0.44	0.73	0.28	0.56
10	0.89	0.22	0.42	0.93	0.21	0.42	0.75	0.27	0.53
11	0.90	0.21	0.40	0.93	0.20	0.40	0.77	0.26	0.50
12	0.91	0.20	0.39	0.94	0.19	0.38	0.79	0.25	0.48
13	0.91	0.19	0.37	0.94	0.19	0.37	0.80	0.24	0.46
14	0.92	0.18	0.36	0.95	0.18	0.35	0.81	0.23	0.45
15	0.92	0.18	0.35	0.95	0.17	0.34	0.82	0.22	0.43
16	0.93	0.17	0.33	0.95	0.17	0.33	0.83	0.21	0.42

Mini-CEX = mini-clinical evaluation exercise; DOPS = directly observed procedural skills; MSF = multi-source feedback; G = generalisability coefficient; SEM = standard error of the mean; 95% CI = 95% confidence interval

given level of reliability. On the basis of the variance component estimates in Table 3, the RCP has recommended a larger sample of mini-CEX assessments than previous investigators, [4] and a correspondingly higher number than is currently recommended for the UK Foundation Programme. This dataset suggests the mini-CEX is subject to significant assessor error – both 'hawk/dove' error (V_a) and error caused by assessors' differing 'average' views of particular trainees (V_{a*t}), and that this source of error outweighs case-specificity. Previous work has also indicated that the mini-CEX is subject to assessor error.[15,16] These estimates, however, should be interpreted cautiously because of the limited sampling of the effects in the regression model. Assessors were not trained for this study, but training may reduce assessor variation and thereby improve the reliability of the mini-CEX and DOPS.

Implications and recommendations

Overall, the three assessment methods have a reasonable balance of utility for informing medical SpR assessment for RITA. However, several important factors need to be addressed in rolling these assessments out.

1 All methods, but the mini-CEX in particular, require an adequate allocation of time and resources.
2 Voluntary participation will not reach a high proportion of trainees and it will be necessary to make these assessments a compulsory requirement for the RITA.
3 An adequate sample of performance is required for all three methods. In particular, the mini-CEX requires sampling across a significant number of assessors in order to fairly represent the view of all assessors. In addition, the mini-CEX should cover a spread of case complexities and settings because a trainee who is assessed only on simple cases or only on outpatient encounters will be disadvantaged. Multi-source feedback should be sought from a balanced sample of raters from all four of the pre-specified groups we used and a consistent mix of genders, because a trainee who is rated only by male consultants, for example, will be disadvantaged.

4 Taking the mean of scores or ratings produces the most reliable overall result. However, it may not be the most sensitive way to pick up doctors with performance difficulties. To illustrate, no doctor in this population obtained a mean aggregate score (across assessors) in the 'unsatisfactory' range. However, some doctors received several individual scores within the 'unsatisfactory' range. The free text comments given on their forms suggest potential specific performance concerns in several of these doctors. It is important for both formative appraisal and for detecting doctors in difficulty that this information is not lost.

Our findings may well be of relevance to the assessment of doctors in other countries.

Further work

Some new observations merit further investigation.

1 The judge error affecting the mini-CEX assessment could have serious implications for the nature and size of the sample required. As this assessment is being used for the UK Foundation Programme, this issue needs further investigation, perhaps by investing in a large, fully-crossed study.
2 The DOPS assessment appears to have very favourable reliability. The data generated by the UK Foundation Programme may allow further investigation into the reliability of the DOPS.
3 The effects of gender and professional group on MSF ratings are very important. The mean ratings given by these different groups vary widely and imply, for example, that a doctor's rating may be more affected by the professional group(s) of the respondents than by the doctor's 'true' performance. It would be valuable to investigate whether this trend exists across different MSF settings.
4 Once the methods are fully implemented in the full cohort of medical SpRs, it will be important to reassess both reliability and pass/fail cut-off values.

Based on our findings the RCP implemented all three of these assessment methods for medical SpRs from October 2006.

Contributors: JRW wrote the final study protocol, ran the study, collected all the results, analysed the qualitative data and wrote the final report. JGMC analysed the quantitative data and advised at every stage, from protocol design to data interpretation. AW wrote the initial study protocol and helped secure funding for the study. WW conceived the study and supervised JRW and AW. All authors were involved in the form design and study design and critically contributed to the final preparation of this article. WW is the guarantor.
Acknowledgements: we are indebted to Dr John Norcini for his advice on study design, Dr Sandy Egan (psychologist) for her advice on study protocol and questionnaire form design, Dr Clive Lewis for his help with study protocol design, the Medical Workforce Unit at the Royal College of Physicians for form design and scanning, and Rachel Oatley for data input. We are grateful to all the consultants who acted as local study co-ordinators.
Funding: JRW was funded full-time for 1 year by a grant from the Sir John Fisher Foundation and equally from funds from the 3 Royal Colleges of Physicians. AW was funded full-time for 3 months by the London Chest Hospital.
Conflicts of interest: none.
Ethical approval: not required.

REFERENCES

1 Wragg A, Wade W, Fuller G *et al*. Assessing the performance of specialist registrars. *Clin Med*, 2003;**3** (2):131–4.
2 General Medical Council. *Good Medical Practice*. London: GMC 2001.
3 Rethans JJ, Norcini JJ, Baron-Maldonado M *et al*. The relationship between competence and performance: implications for assessing practice performance. *Med Educ*, 2002;**36** (10):901–9.
4 Norcini JJ, Blank LL, Duffy FD *et al*. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med*, 2003;**138** (6):476–81.
5 Ramsey PG, Wenrich MD, Carline JD *et al*. Use of peer ratings to evaluate physician performance. *JAMA*, 1993;**269** (13):1655–60.
6 Wenrich MD, Carline JD, Giles LM *et al*. Ratings of the performances of practising internists by hospital-based registered nurses. *Acad Med*, 1993;**68** (9):680–7.
7 Tunbridge M, Dickinson D, Swan P. Outcomes of assessments of registrars in the medical specialties. *Clin Med*, 2004;**4** (1):66–8.
8 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ*, 2005;**330** (7502):1251–3.
9 Frank JR, Langer B. Collaboration, communication, management, and advocacy: teaching surgeons new skills through the CanMEDS Project. *World J Surg*, 2003;**27** (8):972–8; Discussion: 978.
10 Norcini JJ. Current perspectives in assessment: the assessment of performance at work. *Med Educ*, 2005;**39** (9):880–9.
11 Crossley J, Davies H, Humphris G *et al*. Generalisability: a key to unlock professional assessment. *Med Educ*, 2002;**36** (10):972–8.
12 Song S, Jung B. A comparative study of alternative estimators for the unbalanced two-way error component regression model. *J Econom*, 2002;**5** (2):480–93.

13 Crossley J, Eiser C., Davies H. Children and their parents assessing the doctor–patient interaction: evaluating the feasibility and reliability of SHEFFPAT – a rating system for doctors' communication skills. *Med Educ*, 2005;**39** (8):820–8.
14 Crossley J, Russell J, Jolly B *et al.* I'm pickin' up good regressions: the governance of generalisability analyses. *Med Educ* 2007;**41** (10):926–34.
15 Norcini JJ, Blank LL, Arnold GK *et al.* Examiner differences in the mini-CEX. *Adv Health Sci Educ Theory Pract*, 1997;**2** (1):27–33.
16 Holmboe ES, Huot S, Chung J *et al.* Construct validity of the mini-clinical evaluation exercise (mini-CEX). *Acad Med*, 2003;**78** (8):826–30.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix S1.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365.2923.2008.03010.x.

(This link will take you to the article abstract).

Please note: Blackwell Publishing Ltd is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than those pertaining to missing material) should be directed to the corresponding author for the article.

Received 17 October 2006; editorial comments to authors 5 February 2007, 13 June 2007; accepted for publication 26 October 2007