# Learning in practice

## Use of SPRAT for peer review of paediatricians in training

Julian C Archer, John Norcini, Helena A Davies


### Abstract

**Objective** To determine whether a multisource feedback questionnaire, SPRAT (Sheffield peer review assessment tool), is a feasible and reliable assessment method to inform the record of in-training assessment for paediatric senior house officers and specialist registrars.

**Design** Trainees' clinical performance was evaluated using SPRAT sent to clinical colleagues of their choosing. Responses were analysed to determine variables that affected ratings and their measurement characteristics.

**Setting** Three tertiary hospitals and five secondary hospitals across a UK deanery.

**Participants** 112 paediatric senior house officers and middle grades.

**Main outcome measures** 95% confidence intervals for mean ratings; linear and hierarchical regression to explore potential biasing factors; time needed for the process per doctor.

**Results** 20 middle grades and 92 senior house officers were assessed using SPRAT to inform their record of in-training assessment; 921/1120 (82%) of their proposed raters completed a SPRAT form. As a group, specialist registrars (mean 5.22, SD 0.34) scored significantly higher ($t = -4.765$) than did senior house officers (mean 4.81, SD 0.35) ($P < 0.001$). The grade of the doctor accounted for 7.6% of the variation in the mean ratings. The hierarchical regression showed that only 3.4% of the variation in the means could be additionally attributed to three main factors (occupation of rater, length of working relationship, and environment in which the relationship took place) when the doctor's grade was controlled for (significant $F$ change $< 0.001$). 93 (83%) of the doctors in this study would have needed only four raters to achieve a reliable score if the intent was to determine if they were satisfactory. The mean time taken to complete the questionnaire by a rater was six minutes. Just over an hour of administrative time is needed for each doctor.

**Conclusions** SPRAT seems to be a valid way of assessing large numbers of doctors to support quality assurance procedures for training programmes. The feedback from SPRAT can also be used to inform personal development planning and focus quality improvements.


### Introduction

Multisource feedback, or peer review, questionnaires have been studied around the world as a way of assessing multiple components of clinical performance and shown to be feasible and acceptable to doctors.[1] They are also reliable across different settings.[2–5] However, some concerns have been raised about the validity of this approach and the paucity of work done with peer ratings in the United Kingdom.[6]

The Sheffield peer review assessment tool (SPRAT) has been used in the South Yorkshire and South Humberside Deanery to assess all paediatricians in training. SPRAT has already been evaluated as a voluntary appraisal tool for paediatric consultants and found to be reliable.[2 7]

SPRAT has been developed to inform the quality assurance process when assessing doctors' work based performance and has been designed for use as part of a performance assessment programme. SPRAT should also contribute to the quality improvement of doctors. In this paper we discuss SPRAT's implementation and feasibility when used as an assessment method to inform the record of in-training assessment for paediatricians. We also discuss SPRAT's validity and reliability and key areas for further work.

### Methods

#### Questionnaire design and distribution

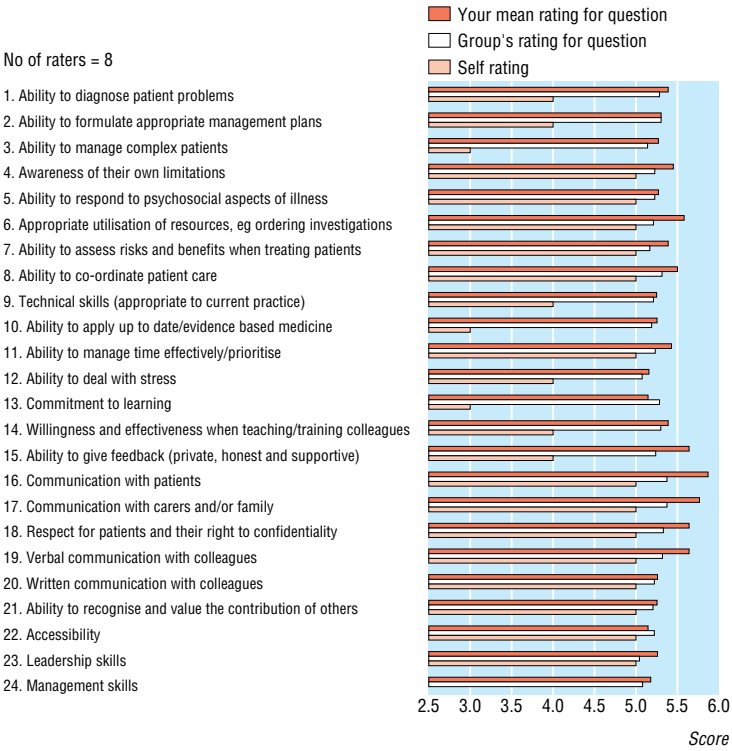
The peer review questionnaire (SPRAT) was designed to assess the components of performance as described in *Good Medical Practice*[8] and by the Royal College of Paediatrics and Child Health.[9] Two authors wrote the questions, which were field tested in two pilot studies at the Sheffield Children's Hospital.[2–7] After modification following feedback, the final form contained 24 questions covering five domains of good medical practice: good clinical care; maintaining good medical practice; teaching and training, assessing and appraising; relationships with patients; and working with colleagues. As it has been mapped explicitly to good medical practice and modified after field testing, SPRAT's content validity has been previously established.

We gathered ratings on a six point scale where 1 = very poor, 4 = satisfactory (the pass mark), and 6 = very good. We provided space for observations and examples.

*This is the abridged version of an article that was posted on bmj.com on 9 May 2005: http://bmj.com/cgi/doi/10.1136/bmj.38447.610451.8F*

Academic Unit of Child Health, Sheffield Children's Hospital, Sheffield S10 2HT
Julian C Archer *clinical research fellow*
Postgraduate Medical Education Department, Sheffield Children's Hospital
Helena A Davies *consultant in medical education*
Foundation for the Advancement of International Medical Education Research, (FAIMER), 3624 Market Street, Philadelphia, PA 19104, USA
John Norcini *president*
Correspondence to: H A Davies h.davies@sheffield.ac.uk

*BMJ* 2005;330:1251–4




Example of feedback chart given to a doctor

The questionnaire was designed to be suitable for completion by raters from any professional background and at any level of training. This approach allowed multisource feedback, increased the feasibility of the method, and made it easy to combine different viewpoints into a single overall evaluation.

We collected data on the clinical setting and the nature of the respondents. Specifically, we recorded the position of the respondent (for example, consultant, nurse), the length of the working relationship with the doctor, and the environment in which the relationship took place (such as outpatients). We also collected data on feasibility, including the amount of time it took to complete the form.

Previous work has shown that raters chosen by people being assessed do not provide significantly different evaluations from those chosen by a third party.[3] We used SPRAT to assess paediatric trainees over an eight month period. We sent them a SPRAT self assessment form with a stamped addressed envelope and asked them to provide the names of raters with whom they worked clinically. We sought 10 nominations, as 8-12 raters are needed to achieve reasonable levels of reliability.[2 3 10–12]

A central administrator contacted the raters in writing and asked them to complete a SPRAT form. The completed forms were returned and collated by the administrator. After processing, and screening by the programme director, we sent copies of the feedback to the doctor and educational supervisor. Feedback consisted of a bar chart showing the doctor's mean score, the group's mean score, and the doctor's self rating score for each question (figure). Comments were typed and fed back to the doctor verbatim. We recorded the amount of administrative time needed to process the forms.

### Study population

The study population consisted of all specialist registrars within the deanery and all senior house officers in a large paediatric trust undergoing mandatory assessment as part of the annual review process for their records of in-training assessment.

### Statistical analysis

*Descriptive analyses*—We calculated frequencies, means, standard deviations, and correlations to describe the participants, the performance of items on the questionnaire, the ratings of the participants, and the feasibility of the method.

*Comparison of groups*—We compared the mean scores achieved by senior house officers and specialist registrars. We also compared full time and part time employment and teaching and non-teaching hospitals.

*Regression*—We used linear regression to explore potential influences on the ratings of the doctors. We did a hierarchical regression controlling for the doctor's grade (senior house officer or specialist registrar), as we accepted that training would affect performance. The three main variables of interest, grouped second, were the length of the working relationship, the working environment (inpatient or outpatient), and the rater's occupation (consultant, middle grade, senior or pre-registration house officer, or nurse).

*Reliability*—To estimate reliability, we calculated a 95% confidence interval for mean ratings on the basis of generalisability theory.[13] We analysed the total scores and estimated variance components for both the trainees and measurement error ("raters nested within trainee") (see bmj.com).

## Results

### Descriptive results

Twenty middle grades and 92 senior house officers were assessed. We sent questionnaires to the 1120 respondents identified. Of these, 921 (82%) completed the forms: 282 (31%) senior or preregistration house officers, 214 (23%) middle grades, 216 (23%) nurses, 186 (20%) consultants, and 13 (1%) others. Ten (1%) raters did not indicate their occupation. The average senior house officer or specialist registrar had eight (range 1-10) completed questionnaires.

The mean ratings of the individual items on the questionnaire at the level of the questionnaire ranged from 4.65 (SD 0.80) to 5.05 (SD 0.82). The lowest ratings were given for "the ability to manage complex patients" and "leadership skills," and the highest ratings were given for "verbal communication with colleagues" and "accessibility." As is typical for ratings forms of this kind, the individual items were very highly intercorrelated, ranging from 0.45 to 0.97. When aggregated to the level of the individual doctor, the mean rating ranged from 3.62 to 5.64 with a mean of 4.89 (SD 0.38).

### Group comparisons

As a group, specialist registrars (mean 5.22, SD 0.34) scored significantly higher than senior house officers (mean 4.81, SD 0.35) ($t = -4.765$, df = 110, $P < 0.001$). We found no statistically significant difference between the performance of doctors working part time (mean

5.00, SD 0.51) and those working full time (mean 4.88, SD 0.38) ($t = -0.582$, df = 99, P = 0.56) or between those working in teaching hospitals (mean 4.88, SD = 0.39) as opposed to district general hospitals (mean 4.94, SD 0.31) ($t = -0.487$, df = 101, P = 0.63).

### Regression

The grade of the doctor accounted for 7.6% of the variation in the mean ratings. The hierarchical regression showed that only 3.4% of the variation in the means could be additionally attributed to the three main factors (occupation of the rater, length of the working relationship, and environment in which the relationship took place) when controlled for the doctor's grade (significant $F$ change $< 0.001$). See bmj.com for details.

### Reliability

Ninety three (83%) of the 112 doctors scored an overall mean of 4.5 or more. When we looked at the 95% confidence levels around the mean score when assessed by varying numbers of raters, we found a 95% confidence interval of ±0.5 when the number of raters was four (see bmj.com). For these 83% of doctors, therefore, only four raters would be needed to achieve a reliable score if the intent was to determine if they were satisfactory.

### Feasibility

Original pack preparation and distribution took 25 minutes per doctor. The mean time taken to complete the questionnaire by a rater was six minutes. The scanning of completed forms took only one second for 10 forms; the verification process and typing of free texts comments took on average 70 seconds per form. Feedback analysis and preparation of reports took an average of 30 minutes.

## Discussion

Multisource feedback has been explored as a way of reliably assessing doctors in the workplace in other countries. We are not aware of published reliability data exploring the use of peer ratings in the United Kingdom.

Evidence for construct validity is provided by the lowest ratings being given to trainees for questions concerned with the management of complex patients and leadership skills, and by specialist registrars scoring significantly higher than senior house officers. SPRAT has been designed not only as a feasible, valid, robust assessment tool to help to inform high stake decisions but also to provide feedback to doctors. This feedback can be used to inform personal development plans. Further work on SPRAT's validity is being done as part of collaborative work with the National Clinical Assessment Service and the Royal College of Paediatrics and Child Health. This will include correlation studies between SPRAT and other instruments, such as mini-CEX,[14] to explore criterion validity. The main sources of bias explored contributed little to the variability in the mean scores. Further studies will look at SPRAT with other cohorts.

SPRAT took just over an hour of administrative time from initially contacting the doctor to the distribution of the doctor's completed feedback profile. Fax and online submission should shorten this time.

### What is already known on this topic

Validated, reliable assessment methods are needed to evaluate doctors in the UK

Multisource feedback has been explored in other countries as a way of assessing traditional and broader competencies, such as professionalism

### What this study adds

Multisource feedback has been evaluated quantitatively for use in the UK

SPRAT seems to be a valid way of reliably informing the record of in-training assessment process

With few raters needed for a robust assessment, SPRAT is a feasible way of assessing behaviours that are traditionally hard to capture

We have not covered the educational impact of SPRAT, but this is planned. Additionally, longitudinal follow-up of doctors assessed using multisource feedback such as SPRAT will allow determination of predictive validity, currently an unexplored aspect of workplace based assessment.

SPRAT represents the first major published work on multisource feedback in the UK. It is reliable, feasible, and practical to instigate in the NHS.

We thank Jean Russell, computer officer-statistician, University of Sheffield, for her support and advice. We also acknowledge the support of Sarah Thomas, postgraduate dean for the South Yorkshire and South Humberside Deanery.

Contributors: See bmj.com

Funding: JCA's research fellowship is funded by cooperation between the Academic Unit of Child Health, University of Sheffield, and Bassetlaw District General Hospital, Worksop.

Competing interests: None declared.

Ethics approval: Not sought. SPRAT was implemented as part of the assessment programme in the South Yorkshire and South Humberside Deanery.

1 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. *Acad Med* 2002;77(10 suppl):S64-6.
2 Archer JC, Davies HA. Clinical management. Where medicine meets management: on reflection. *Health Serv J* 2004;114(5903):26-7.
3 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655-60.
4 Hall W, Violato C, Lewkonia R, Lockyer J, Fidler H, Toews J, et al. Assessment of physician performance in Alberta: the physician achievement review. *CMAJ* 1999;161:52-7.
5 Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. *J Gen Intern Med* 1999;14:551-4.
6 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004;328:1240.
7 Archer JC, Davies HA. *Sheffield peer review assessment tool for consultants (SPRAT): screening for poorly performing doctors*. Bern, Switzerland: Association of Medical Education of Europe, 2003.
8 General Medical Council. *Good medical practice*. London: GMC, 2001.
9 Royal College of Paediatrics and Child Health. *Good medical practice in paediatrics and child health: duties and responsibilities of paediatricians*. London: Royal College of Paediatrics and Child Health, 2002.
10 Violato C, Marini A, Towes J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers and patients to assess physicians. *Acad Med* 1997;72(suppl 1):S82-4.
11 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ* 2003;326:546-8.
12 Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol* 2002;99:647-51.
13 Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. *Educ Psychol Measurem* 2004;64:391-418.
14 Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004;79:16-22.

*(Accepted 1 April 2005)*

doi 10.1136/bmj.38447.610451.8F