

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

Several regulatory processes in the United Kingdom aim to increase the assessment of doctors in training and practice.^{1,2} Robust, valid, feasible, and acceptable measures of competence are essential to support these efforts. We aimed to explore multisource feedback as one potential response to these needs.

Multisource feedback, or peer review, questionnaires have been studied around the world as a way of assessing multiple

components of clinical performance.³⁻¹⁴ They have been shown to be feasible and acceptable to doctors,⁴ which is fundamental to a tool's success.¹⁵ They are also reliable across different settings.^{10, 12-14} However, some concerns have been raised about the validity of this approach and the paucity of work done with peer ratings in the UK.¹⁶

Building on previous work on multisource feedback done in the UK,^{10, 11} the Sheffield peer review assessment tool (SPRAT) has been implemented in the South Yorkshire and South Humberside Deanery to assess all paediatricians in training. The feedback from SPRAT is used to inform the record of in-training assessment (RITA). Previous published work evaluated the initial development and application of SPRAT as a voluntary appraisal tool for paediatric consultants.^{10, 11} The tool was found to be reliable, but some questions were infrequently answered as they were poorly phrased. This led to subsequent remodelling of the instrument before it was used further.

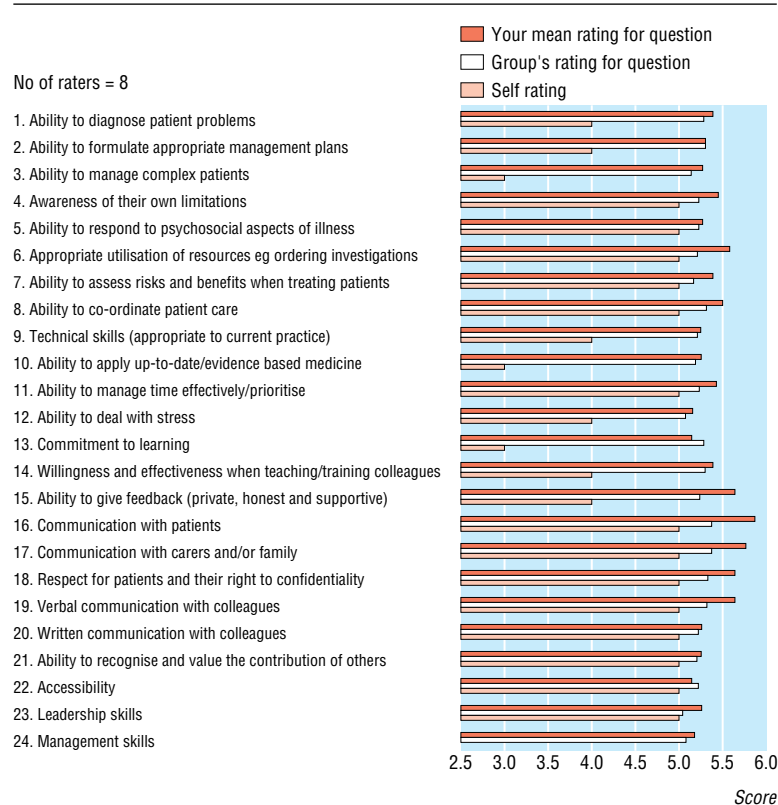
SPRAT has been developed to inform the quality assurance process when assessing doctors' work based performance. It has been designed for use not in isolation but as part of a performance assessment programme. SPRAT should also contribute to the quality improvement of doctors, but work in this area is at an early stage. 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 in training. We also discuss SPRAT's validity and reliability and identify 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 the document *Good Medical Practice (GMP)*¹⁷ and made relevant to this clinical context by the Royal College of Paediatrics and Child Health.¹⁸ Two of the authors (JCA and HAD) wrote the questions, which were field tested in two pilot studies at the Sheffield Children's Hospital^{10, 11} following accepted good practice.¹⁹ The wording of some questions was modified as a result of qualitative feedback received. An item total correlation analysis did not lead to the removal of any items; the lowest correlation was 0.45. The final form contains 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, 2 = poor, 3 = needs development, 4 = satisfactory (the pass mark), 5 = good, and 6 = very good. Respondents who did not



Example of feedback chart given to a doctor

observe a specific behaviour could indicate that they were “unable to comment.” We provided space for observations and examples.

The questionnaire was designed to be suitable for completion by raters from any professional background and at any level of training (such as a senior sister or a preregistration house officer). In contrast to an approach with different forms for different occupational groups or the use of a single occupational group,^{12, 13} this approach was consistent with the desire for multisource feedback, increased the feasibility of the method, and made it easy to combine different viewpoints into a single overall evaluation.

In addition to the ratings, we collected data on the clinical setting and the nature of the respondents to determine their effect on the results. 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) for each questionnaire. We also collected data bearing 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.¹² 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.^{6, 7, 9-12}

A central administrative office contacted the raters in writing and asked them to complete a SPRAT form. The completed forms were returned to the administrative office and then scanned, verified, and collated into a spreadsheet.

After all the data had been processed, we sent copies of the feedback to the doctor and his or her educational supervisor.

The programme director screened the feedback before it was posted to identify any particular areas of concern. The feedback consisted of a bar chart showing the doctor's mean for each question compared with the group's mean for that question, as well as the doctor's self rating score (figure). We also gave doctors their overall mean for the questionnaire compared with the group's. Comments were typed and fed back to the doctor verbatim. As a measure of feasibility, 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 who were being assessed as part of the annual review process to inform their records of in-training assessment. Participation was mandatory as part of this assessment process.

Statistical analysis

We used SPSS version 11.0 to analyse the data. We removed questions marked “unable to comment” before analysis.

Descriptive analyses—We calculated frequencies, means, standard deviations, and Pearson product moment 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 used *t* tests to compare the mean scores achieved by senior house officers and specialist registrars. We also used *t* tests to compare 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 preregistration house officer, or nurse). We divided the length of working relationship into quarters.

Reliability—To estimate reliability, we calculated a 95% confidence interval for mean ratings on the basis of generalisability theory.²⁰ We used the VARCOMP procedure in SPSS to analyse the total scores and estimate variance components for both the trainees and measurement error ("raters nested within trainee"). The square root of the variance component for measurement error is the standard error of measurement; we calculated this for 1-10 raters ($\sqrt{\text{error}/\text{number of raters}}$). The 95% confidence intervals are equal to the standard error of measurement multiplied by 1.96 and are added to and subtracted from a mean rating.

Results

Descriptive results

Twenty middle grades and 92 senior house officers were assessed to inform their records of in-training assessment. 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, however, 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). One doctor fell short of the overall pass mark.

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). (Grade of doctor (senior house officer or specialist registrar), unstandardised $\beta = 0.41$, standardised $\beta = 0.28$, $P < 0.001$. Working environment (inpatient or outpatient), unstandardised $\beta = -0.13$, standardised $\beta = -0.10$, $P < 0.001$. Occupation of rater: consultant, unstandardised $\beta = -0.24$, standardised $\beta = -0.19$, $P < 0.05$; middle grade, unstandardised $\beta = -0.22$, standardised $\beta = -0.19$,

Confidence levels for mean score based on 1-10 raters

No of raters	95% confidence interval
1	± 1.0
2	± 0.7
3	± 0.6
4	± 0.5
5	± 0.4
6	± 0.4
7	± 0.4
8	± 0.3
9	± 0.3
10	± 0.3

$P < 0.05$; senior house officer, unstandardised $\beta = -0.13$, standardised $\beta = -0.15$, $P > 0.05$; nurse, unstandardised $\beta = -0.07$, standardised $\beta = -0.08$, $P > 0.05$. Length of working relationship, unstandardised $\beta = 0.04$, standardised $\beta = 0.10$, $P < 0.001$).

Reliability

The table summarises the 95% confidence levels around the mean score when assessed by varying numbers of raters. Ninety three of the 112 doctors in this study scored an overall mean of 4.5 or more. 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. In other words, we can be 95% confident that a doctor scoring 4.5 or above has been correctly passed, as the lowest score likely on retesting is 4.0 (still the expected standard). Doctors achieving a mean score nearer to or below the pass mark would need additional raters to be confident of their correct placement around the pass mark.

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 UK. Concerns about the paucity of research in this field have recently been highlighted.¹⁶

In this study, 112 paediatric trainees were assessed by their clinical colleagues (fellow trainees, consultants, nurses, and other health professionals) using a generic questionnaire, SPRAT. The response rate was 82%.

The fact that the lowest ratings given to the trainees were those concerned with the management of complex patients and leadership skills provides evidence of construct validity. Further evidence for construct validity is provided by specialist registrars scoring significantly higher than senior house officers.

A regression analysis showed that when we controlled for the differences between the grades of the doctor only 3.4% of the variation in the means could be attributed to the length of the working relationship, the occupation of the rater, and the working environment. This can be ignored. Of the 112 doctors in the study, 93 (83%) scored mean ratings above 4.5. Using 95% confidence intervals (table), we can say that only four raters would

need to assess these doctors to be confident that they had appropriately passed. If there are concerns about a trainee or if the trainee is borderline, further raters would be needed. The raters took six minutes on average to complete a SPRAT form.

SPRAT is a feasible tool for use in the NHS. It 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. The tool's development and validation is in line with good practice guidance laid out in the Postgraduate Medical Education Training Board's principles for assessment document.²

Previous work has established SPRAT's content validity.^{10 11} SPRAT's validity will need to be explored further. This is being done as part of collaborative work with the National Clinical Assessment Service and the Royal College of Paediatrics and Child Health. This work will include correlation studies between SPRAT and other instruments, such as mini-CEX,²¹ to explore criterion validity.

The main sources of bias that were explored in this paper contributed little to the variability in the mean scores. We must acknowledge that others factors might be relevant but not identified here and that all these biases may not generalise as such in further studies. SPRAT must, and will, therefore be scrutinised with other cohorts.

SPRAT took just over an hour of administrative time for each doctor for the whole process from initially contacting the doctor to the distribution of the doctor's completed feedback profile. Fax and online submission are being installed, and we hope that these will further increase SPRAT's feasibility. We have not covered the educational impact of SPRAT in this paper. Some evidence shows that receiving feedback can have a positive effect.²² Future work is planned to explore the educational impact on trainees and on those who are identified as having problems. If the NHS is to invest considerable resources in terms of both time and money into ensuring that all doctors undergo robust workplace based assessment, we must evaluate the potential for behavioural change in response to feedback. Provision of structured remediation frameworks will facilitate this, but several models may need to be explored.

Additionally, longitudinal follow-up of doctors assessed using multisource feedback such as SPRAT will allow determination of predictive validity, a vital but currently unexplored aspect of workplace based assessment.

SPRAT represents the first major published work on multisource feedback in the UK. SPRAT is reliable at feasible levels and is practical to instigate in the NHS. It seems to be a valid way of assessing large numbers of doctors to support quality assurance procedures for training programmes. In line with good assessment practice, SPRAT can be used to provide structured feedback across the domains of good medical practice and can thus be used to focus quality improvements.

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: JCA and HAD designed the original study and oversaw its implementation. JCA analysed the data with assistance from JN. JCA wrote the paper with assistance from JN and HAD. HAD is the guarantor.

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.

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

- 1 Department of Health. *A guide to specialist registrar training*. London: NHS Executive, 1998.
- 2 Southgate L, Grant J, Working Group. *Principles and standards for an assessment system for postgraduate medical training*. London: Postgraduate Medical Education and Training Board, 2004.
- 3 Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. *Acad Med* 2004;79(10 suppl):S5-8.
- 4 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.
- 5 Whitehouse A, Walzman M, Wall D. Pilot study of 360 degree assessment of personal skills to inform record of in-training assessments for senior house officers. *Hosp Med* 2002;63:172-5.
- 6 Violato C, Marini A, Tows 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.
- 7 Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ* 2003;326:546-8.
- 8 Sargeant JM, Mann KV, Ferrier SN, Langille DB, Muirhead PD, Hayes VM, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. *Acad Med* 2003;78(10 suppl):S42-4.
- 9 Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol* 2002;99:647-51.
- 10 Archer JC, Davies HA. Clinical management. Where medicine meets management: on reflection. *Health Serv J* 2004;114(5903):26-7.
- 11 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.
- 12 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.
- 13 Hall W, Violato C, Lewkonja R, Lockyer J, Fidler H, Toews J, et al. Assessment of physician performance in Alberta: the physician achievement review. *CMAJ* 1999;161:52-7.
- 14 Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. *J Gen Intern Med* 1999;14:551-4.
- 15 Finucane PM, Barron SR, Davies HA, Hadfield-Jones RS, Kaigas TM. Towards an acceptance of performance assessment. *Med Educ* 2002;36:959-64.
- 16 Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ* 2004;328:1240.
- 17 General Medical Council. *Good medical practice*. London: GMC, 2001.
- 18 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.
- 19 Norcini J. Peer assessment of competence. *Med Educ* 2003;37:539-43.
- 20 Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. *Educ Psychol Measur* 2004;64:391-418.
- 21 Holmboe ES. Faculty and the observation of trainees' clinical skills: problems and opportunities. *Acad Med* 2004;79:16-22.
- 22 Ringsted C, Henriksen AH, Skaarup AM, Van der Vleuten CPM. Educational impact of in-training assessment (ITA) in postgraduate medical education: a qualitative study of an ITA programme in actual practice. *Med Educ* 2004;38:767-77. (Accepted 1 April 2005)

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