User perceptions of multi-source feedback tools for junior doctors: usability, usefulness and validity

1. Introduction

Multi-source feedback (MSF) refers to feedback on job performance given by people with whom the recipient works. In recent years it has developed from origins in the commercial and management sectors to become a significant element of medical education and revalidation[1 2]. MSF typically involves the completion of a questionnaire tool by a number of colleagues ('raters'), whose responses are summarised to identify areas of strength or weakness. Scores can be compared with a criterion of performance, or with the population scores of the recipient's peer group.

Since 2005, MSF has been a core feature of a learning portfolio completed by doctors in the UK during the first two years of their generic postgraduate training (the Foundation Programme). As well as MSF, this portfolio contains reflective accounts, reports from supervisors, and a number of other workplace assessments – a mini clinical examination, direct observation of clinical procedures, and case-based discussion[3]. Each of these is completed by a clinical colleague who rates different elements of practice. The portfolio specifies how many of each must be completed during each Foundation Programme year.

Two MSF tools are used in England, Wales and Northern Ireland: the mini-Peer Assessment Tool (mini-PAT[4]) and Team Assessment of Behaviour (TAB[5]), while a third is used in Scotland (there is no work on this tool currently published). TAB was developed to provide feedback to doctors working in obstetrics and gynaecology[6], while mini-PAT was developed from a feedback tool for paediatricians[7]. Data has been published on the reliability and validity of both[4 5], and on inter-professional differences in scores for TAB[8].

Little work has looked at the perceptions of people who actually use MSF tools in practice. Some data on TAB users' perceptions has been published, showing trainees, their supervisors and raters having generally positive attitudes[9]. Elsewhere, work on the Practitioner Achievement Review (PAR) in Canada found that recipients' attitudes to feedback varied with factors such as perceptions of its credibility[10], and their own mood and ability to reflect[11]. A study in Scotland[12] found that raters, recipients and supervisors had positive opinions of a tool's ease of use and usefulness, but had concerns whether raters would have sufficient knowledge to give accurate feedback. Similar findings were reported in a study with medical students[13]. A survey of general practitioners' opinions of different assessments found that MSF was felt to be less acceptable and feasible than patient feedback, audit and significant event analysis[14].

These attitudes are important: it is essential that people who use a tool or system accept it, or it will fail (an extension of the principles of the Technology Acceptance Model[15 16]). To be accepted, an MSF system needs to be both usable and useful – it must be workable in practice, and have intended positive outcomes. With MSF, raters must feel a tool allows them to give necessary feedback which will benefit recipients, or they may not take the time to consider their responses. Recipients should feel it is worthwhile and valid, or they may not respond to it. Where feedback is delivered by a third party such as a supervisor, they must also see value, or it may not be delivered effectively.

1.1 The current study

The Northern Deanery, responsible for postgraduate medical and dental education across the North East of England (and Cumbria to the west), supported both mini-PAT and TAB in the first year of the Foundation Programme (August 2005 to July 2006).

Five hospital trusts used mini-PAT, while four used TAB. This provided a unique opportunity to compare the attitudes of users of the two tools within a single region. At the time of the study both TAB and mini-PAT were pen-and-paper questionnaires, and differed in two main ways:

- Length of form: mini-PAT had 16 items, mapped in detail to Foundation Programme competencies including areas of clinical practice, while TAB had four items describing broader interpersonal areas of practice (each with examples). Mini-PAT additionally had a single item with a 'Yes/No' response to highlight concerns about probity or health, and seven items for demographics and professional details. TAB had just one additional item to record the rater's job. TAB was presented on one side of paper, mini-PAT on two.

- Primary mode of feedback: both tools contained numerical scale and free-text feedback, but the focus of each was different. TAB responses were mainly textual, with large free-text areas and a three-category scale response for each item, whereas mini-PAT responses were on a six-point numerical scale for each of the 16 items, and a single free-text area prompting for positive and negative comments on the reverse.

The tools therefore represent different approaches to feedback tool design.

For both tools, raters were chosen by trainees. TAB questionnaires were distributed directly by trainees, while mini-PAT was distributed centrally following nomination of raters by trainees.

The study reported here looked at two main questions:

- did the opinions of users of the two tools differ?

- did the opinions of the different user groups differ?

Opinions relating to the usability, usefulness and validity of the tools were obtained by questionnaire. The areas of interest had been identified in pilot work carried the previous year, when a different learning portfolio and MSF tool had been in use in the region as part of a Foundation Programme pilot[17]. The pilot work had consisted of a questionnaire study across two hospital trusts, and telephone interviews with trainees and people providing feedback at one of those trusts[18].

The delivery of both TAB and mini-PAT has changed since the time of the study, with both now completed and delivered electronically in most areas of the UK. However, both are largely unchanged in content (mini-PAT now includes two distinct text areas in place of the additional seven items recording personal information). Current versions can be found in the Foundation Programme learning portfolio[19].

2. Method

Ethical approval for the study was obtained from the Hull and East Riding NHS Research Ethics Committee.

2.1 Participants

Participants in the study were drawn from three populations:

- 'Trainees': Doctors in the first year of the Foundation Programme, which involved three four-month placements in different specialities;

- 'Raters': The colleagues who were asked to give feedback. These could be members of any professional group, clinical or non-clinical, with whom the trainees had worked or were currently working;

- 'Supervisors': The trainees' educational supervisors, who were named consultant doctors with responsibility for the trainees for that year, and did not necessarily work clinically with them. In the context of MSF, their role was to mediate the feedback, receiving an aggregated summary which they would deliver in a face-to-face meeting, identifying strengths and weaknesses and an action plan if necessary. Educational supervisors could also act as raters for other trainees with whom they worked.

2.2 Materials

Questionnaires were developed from those used in pilot work the previous year. Questions covered different areas of the feedback tools' usability, usefulness and validity with responses on a five point Likert scale (‘Strongly disagree’ to ‘Strongly agree’). Questions were designed to be treated as single-item scales, rather than as aggregated constructs. Different questionnaires were developed to reflect the different relationships of trainees, raters, and supervisors to the tool. Some items were worded identically while others reflected different roles (the actual wording of each question is included in the results).

To ensure respondents focused on the appropriate tool, they were asked which they had used, with images of both on the front page of the questionnaire.

2.3 Distribution of questionnaires

Questionnaires were intended for the entire population of Foundation Programme trainees in the Northern Deanery (n=510), educational supervisors (n~=364), and raters (n~=1833). The actual sizes of supervisor and rater populations were unknown, as no central list existed. For eight of the nine trusts involved, rater and supervisor names were obtained from the local education centres in hospitals, but for the ninth details were not available. A batch of 200 rater questionnaires and 78 supervisor questionnaires was sent to that trust for distribution. It is not known how many of these reached intended recipients.

Hard copies of the questionnaires, presented as a booklet of four pages, were sent to respondents in January 2006, with two reminders sent after four and eight weeks.

3. Results

3.1 Response rate

Overall, 45% of all questionnaires produced were returned. For trainees the rate was 53% (249/467 – 38 were returned undelivered), for supervisors 44% (161/364) and for raters 45% (829/1833). For each tool (deriving the numerator from the Trust indicated by respondents, and the denominator from the number sent to each Trust), the overall response rates are, for mini-PAT: trainees 174/333 (52%), raters 548/1249 (44%), supervisors 105/230 (46%) and for TAB: trainees 74/134 (55%), raters 238/584 (41%), supervisors 43/134 (32%).

These are good response rates for a study of this sort. However they may be deflated by other factors, including the lack of direct distribution for one trust and some details of

educational supervisors being out of date. Some trainees and supervisors had not completed the MSF cycle during the timescale of the questionnaire and so could not complete the questionnaire. Effective response rates can therefore be inferred to be even higher.

3.2 Analysis

The analysis looked for differences between the responses of people who had used TAB or mini-PAT and, where comparable questions were asked, between the different groups using them (trainees, raters or supervisors).

Respondents who did not clearly indicate which tool they had used were excluded from analysis. Other responses with more than 25% of items left blank were also excluded from all analysis. This left 679 respondents who had indicated they had used mini-PAT (124 trainees, 83 supervisors and 472 raters) and 222 who had used TAB (48 trainees, 19 supervisors and 155 raters).

Analysis was conducted in SPSS v15. Results from analysis of variance (ANOVA) are reported using the Type 1 sum of squares to compensate for the unequal sample sizes[20]. Nevertheless, the observed power of some analyses is low (statistics generated by SPSS range from 0.052 to 0.994). Statistical power relates to the likelihood of a Type II error – a false negative. An under-powered analysis risks significant differences below a certain size being 'missed', especially where effects are small, as they are here (Cohen's f statistic for ANOVA[21], where effects are significant, ranges from 0.07 to 0.27). To minimise the risk of Type II error in these circumstances, an alpha-correction for multiple tests is not applied[22].

3.3 Comparison of tools

Compared items fall into four areas: general opinions, perceived ease of use, perceived usefulness and potential threats to validity. Descriptive statistics for all items analysed are provided in tables – table 1 gives figures for overall views and items addressing workload and ease of use.

Where an identical question was asked of all three groups, significant differences from a two-way analysis of variance (ANOVA) is reported (independent variables tool and group, both between subjects). For other items, one-way ANOVA is reported (for tool).

Table 1. Descriptive statistics for overall and workload items. Items included in two-way analyses are shaded.							
		Mini-PAT			**TAB**		
Question	**Group**	**Mean**	**SD**	**CI**	**Mean**	**SD**	**CI**
Overall views							
Multi-source feedback is a good idea in principle	trainee	4.11	.87	3.96-4.27	4.00	1.05	3.69-4.31
Multi-source feedback is a good idea in principle	supervisor	3.60	1.20	3.34-3.86	3.89	.81	3.50-4.28
Multi-source feedback is a good idea in principle	rater	3.97	.86	3.89-4.05	4.11	.79	3.99-4.23
Overall, how positive or negative has the feedback been that you have received through this form *('Entirely negative' to 'entirely positive')*	trainee	4.16	.74	4.03-4.29	4.46	.59	4.28-4.63
How appropriate did you find the level of detail or focus of the questions?	trainee	3.12	.74	2.99-3.26	3.46	.91	3.19-3.73
How appropriate did you find the level of detail or focus of the questions?	supervisor	3.03	1.01	2.80-3.25	2.89	.88	2.47-3.32
How appropriate did you find the level of detail or focus of the questions?	rater	3.43	.77	3.36-3.50	3.51	.81	3.38-3.64
Workload and ease of use							
The multi-source feedback form was easy to complete	rater	3.89	0.85	3.81-3.97	3.98	0.76	3.86-4.10
How easy was it to summarise the multi-source feedback for the trainee? *('Simple'-'Difficult', reversed scoring)*	supervisor	3.64	1	3.42-3.86	4.11	0.94	3.65-4.56
How much time did you spend completing the form?	rater	13.64	8.95	12.83-14.46	13.47	8.29	12.14-14.8
How long did you spend preparing the feedback to give to each trainee?	supervisor	15.59	9.57	13.47-17.71	17.89	13.78	11.26-24.53
Number of forms completed this year?	rater	2.48	2.23	2.27-2.69	2.16	1.4	1.93-2.39
Number of FP trainees this year?	supervisor	2.90	1.78	2.52-3.29	2.83	1.95	1.86-3.80

3.4 General opinion of MSF

Respondents generally responded positively to the statement ‘MSF is a good idea in principle’. There was no significant difference between the tools and an overall mean of 3.98 on the five point scale (SD=0.90, CI=3.92-4.04). A significant difference between groups ($F(2, 891)=7.701$, $p<0.001$, $f=0.13$) indicated that supervisors’ opinions were less positive than raters’ and trainees’.

Overall opinion correlated moderately with some items related to ease of use and usefulness (ranging from r=0.239 for 'Could identify a doctor in difficulty' to r=0.457 for 'Could change behaviour or attitudes'). However, there was only a low correlation (r=0.160) with trainees' reports of their feedback being positive or negative. TAB feedback was felt to be more positive than that from mini-PAT ($F(1,167)=5.879$, $p<0.05$, $f=0.19$).

Previous experience of MSF may affect attitudes, so the overall opinions of those who had previous experience were compared with those who had not. Supervisors were asked whether they had received MSF themselves in the past. A significant difference on the ‘good idea in principle’ item ($t(98)=3.628$, $p<0.01$, Cohen’s d=0.73) was found between those who said they had received feedback (n=41; mean=4.12, SD=0.75, CI=3.88-4.36) and those who had not (n=59; mean=3.32, SD=1.26, CI=2.99-3.65). Raters were asked whether they had completed MSF forms in previous years, and no significant difference was found between the overall opinion of those who reported having completed a form before (n=161) and those who had not (n=461).

3.5 Workload and ease of use

No significant difference between tools was found in the reported ease of using the forms, or the time commitment reported by raters and supervisors. Raters estimated they took a mean 13.59 minutes (SD=8.8, CI=12.90-14.29, range 1 to 60) to complete feedback, and supervisors a mean 15.71 minutes (SD=10.59, CI=13.63-17.79, range 4 to 60) to prepare to deliver it. Raters estimated that they completed tools for a rounded mean of two trainees (a range of 1 to 25, with 90% completing four or fewer) and supervisors estimated delivering feedback to three (range 1 to 9, with 88% having five or fewer trainees).

There was no significant difference between the tools in responses to a question about the appropriateness of their detail. There was however a difference between groups ($F(2,884)=16.175$, $p<0.001$, $f=0.19$), with raters being most positive, followed by trainees, then supervisors.

3.6 Perceived usefulness

Table 2 has descriptive statistics for items addressing perceived usefulness. Perceptions of the tools' utility for identifying a doctor in difficulty were low, with an overall mean of 2.99 (SD=1.10, CI=2.92-3.07). However, significant effects were found both for tool ($F(1,885)=4.092$, $p<0.05$, $f=0.07$), with TAB scoring slightly more highly, and for group ($F(2,885)=11.768$, $p<0.001$, $f=0.16$), with higher scores from raters, lower from trainees.

Table 2. Descriptive statistics for 'usefulness' items. Items included in two-way analyses are shaded.							
		Mini-PAT			**TAB**		
Question	**Group**	**Mean**	**SD**	**CI**	**Mean**	**SD**	**CI**
Usefulness							
I think the feedback provided by this form would successfully identify a doctor in difficulty	trainee	2.56	1.19	2.35-2.77	2.89	1.18	2.55-3.24
The feedback provided by this form would successfully identify a doctor in difficulty	supervisor	2.90	1.15	2.65-3.16	3.00	1.05	2.49-3.51
The feedback provided by this form would successfully identify a doctor in difficulty	rater	3.06	1.07	2.97-3.16	3.21	.98	3.05-3.37
Multi-source feedback will lead to positive changes in junior doctors' behaviour and/or attitudes	supervisor	2.84	1.05	2.61-3.07	3.53	.77	3.15-3.9
Multi-source feedback will lead to positive changes in junior doctors' behaviour and/or attitudes	rater	3.31	.96	3.23-3.4	3.44	.96	3.29-3.59
I have changed/will change: Relationships with patients	trainee	2.31	1.10	2.11-2.52	2.52	1.17	2.17-2.88
I have changed/will change: Working with colleagues	trainee	2.49	1.13	2.28-2.7	2.55	1.25	2.17-2.92
I have changed/will change: Clinical care	trainee	2.55	1.08	2.35-2.75	2.77	1.10	2.44-3.11
I have changed/will change: Medical knowledge	trainee	2.66	1.16	2.45-2.88	2.91	1.13	2.58-3.25
I have changed/will change: Teaching and training skills	trainee	2.61	1.13	2.41-2.82	2.64	1.11	2.31-2.98
I have changed/will change: Attitude and approach to job	trainee	2.47	1.15	2.26-2.68	2.62	1.30	2.23-3.01
I have changed/will change: Professional skills (record-keeping, time management etc)	trainee	2.53	1.15	2.32-2.75	2.73	1.32	2.33-3.13
The multi-source feedback I received has been useful and valuable to my learning so far this year	trainee	2.91	.98	2.74-3.09	3.17	1.00	2.88-3.46
How useful do you think the feedback [you received] from this form was in each of these areas...							
...relationships with patients?	trainee	3.48	0.89	3.32-3.64	3.8	1.11	3.48-4.13
	supervisor	3.37	1.21	3.1-3.64	3.72	0.83	3.31-4.13
...working with colleagues?	trainee	3.56	.97	3.38-3.73	3.98	.92	3.71-4.25
	supervisor	3.57	1.16	3.31-3.82	3.72	.75	3.35-4.1
...clinical care?	trainee	3.41	.96	3.24-3.58	3.45	1.08	3.13-3.76
	supervisor	3.23	1.14	2.98-3.48	3.5	.62	3.19-3.81
...medical knowledge?	trainee	3.11	0.97	2.93-3.28	3	1.21	2.64-3.36
	supervisor	2.99	1.17	2.73-3.25	3	0.73	2.61-3.39
...teaching and training skills?	trainee	2.95	0.96	2.77-3.12	3.18	1.17	2.83-3.54
	supervisor	2.91	1.22	2.63-3.19	2.94	1.11	2.39-3.5

Table 2. Descriptive statistics for 'usefulness' items. Items included in two-way analyses are shaded.							
		Mini-PAT			**TAB**		
Question	**Group**	**Mean**	**SD**	**CI**	**Mean**	**SD**	**CI**
...attitude and approach to job?	trainee	3.57	.96	3.4-3.74	3.85	.93	3.58-4.12
	supervisor	3.36	1.17	3.1-3.62	3.89	.83	3.48-4.3
...professional skills (record-keeping, time management etc)?	trainee	3.4	.98	3.23-3.58	3.66	1.10	3.33-3.99
	supervisor	3.3	1.13	3.06-3.55	3.89	.68	3.55-4.23

On the question of whether feedback could change practice, there was a significant overall effect of tool on rater and supervisor responses ($F(1,717)=5.927$, $p<0.05$, $f=0.08$), with TAB felt to be more useful than mini-PAT. Raters were more positive than supervisors ($F(1,717)=12.298$, $p<0.001$, $f=0.13$). A significant interaction ($F(1,721)=4.496$, $p<0.05$, $f=0.08$) indicated that within users of each tool, supervisors scored TAB more highly than raters, while raters scored mini-PAT more highly than supervisors.

Trainees were asked in more detail whether they would change their behaviour on a number of dimensions (relationship with patients, working with colleagues, clinical care, medical knowledge, teaching, attitudes and professional skills) which were derived from the General Medical Council's *Good Medical Practice*[23] and pilot work[11]. No significant differences between tools were found, and no means were higher than the mid-point of the scale (although the confidence interval for TAB does straddle the mid-point), indicating the expected influence of the feedback was low. The correlation between intention to change in any area and the perceived positivity or negativity of feedback was extremely low ($r<0.1$ for all items).

There was a neutral mean and no significant difference between the tools on trainees' responses to whether the feedback had been 'useful and valuable to their learning'

(mean=2.98, SD=0.99, CI=2.83-3.13). However, only 31% of trainees agreed or strongly agreed with the statement.

3.7 Usefulness in different areas of practice

Trainees and supervisors were asked how useful they felt the tool was for feedback in different areas of practice. There were no significant differences between groups, but four between tools (table 3), which indicated that both trainees and supervisors using TAB felt it to be more useful on items related to communication and professionalism.

Table 3. Significant ANOVA results, comparing TAB and mini-PAT users' responses to the question 'How useful was the tool in giving feedback on each of the following...'

Item	**ANOVA result**
...relationships with patients?	$F(1,263)=5.817$, $p<0.05$, $f=0.15$
...working with colleagues?	$F(1,264)=5.755$, $p<0.05$, $f=0.15$
...attitude?	$F(1,264)=6.736$, $p<0.05$, $f=0.16$
...professional skills?	$F(1,262)=5.857$, $p<0.05$, $f=0.15$

3.7.1 Response format

Table 4 gives figures for items relating to response format and validity. Both TAB and mini-PAT incorporated scale and text modes, though their designs emphasised one over the other (mini-PAT being dominantly numerical, with questions on a six point scale; TAB being dominantly text-based with a three point scale). An effect of tool on the usefulness of text feedback ($F(1,884)=9.861$, $p<0.005$, $f=0.11$) indicated that users of TAB found text feedback more useful than those of mini-PAT. An effect of group was found for numerical data ($F(2,875)=32.455$, $p<0.001$, $f=0.27$), with trainees and supervisors rating it lower than raters.

Table 4. Descriptive statistics for 'response format' and 'validity' items. Items included in two-way analyses are shaded.		Mini-PAT			TAB		
Question	**Group**	**Mean**	**SD**	**CI**	**Mean**	**SD**	**CI**
Response format							
How useful to you was the feedback in the form of a numerical rating scale?	trainee	2.85	.98	2.68-3.03	3.04	.98	2.75-3.34
How useful were the rating scales (tick-boxes) for providing necessary and appropriate feedback to the trainee?	supervisor	3.05	1.04	2.82-3.28	3.26	.99	2.79-3.74
How useful were the rating scales (tick-boxes) for giving the feedback you wanted to?	rater	3.50	.91	3.41-3.58	3.55	.85	3.41-3.69
How useful to you was the feedback in the form of free text comments?	trainee	3.82	1.09	3.63-4.02	3.89	.99	3.60-4.19
How useful were the text comments for providing the necessary and appropriate feedback to the trainee?	supervisor	3.43	1.16	3.16-3.69	4.00	1.00	3.52-4.48
How useful were the spaces for writing a comment for giving the feedback you wanted to?	rater	3.65	.87	3.57-3.73	3.86	.79	3.74-3.99
Validity							
I think the feedback I was given on this form was reliable and trustworthy	trainee	3.40	.86	3.24-3.55	3.47	.88	3.21-3.73
I think the feedback returned on the forms was reliable and trustworthy	supervisor	3.16	1.03	2.93-3.38	3.21	.92	2.77-3.65
I am concerned some ratings or comments were not based on actual experience of my work	trainee	2.69	1.16	2.48-2.9	3.15	1.24	2.79-3.51
I had sufficient experience of the doctor's work to give accurate ratings	rater	3.78	.90	3.7-3.86	3.79	1.01	3.63-3.95
I know they have experience of my work	trainee	4.49	0.75	4.36-4.63	4.49	0.66	4.3-4.68
I get on with them as a person	trainee	3.69	0.89	3.53-3.84	3.78	1.03	3.48-4.09
I expected to get positive feedback from them	trainee	3.16	0.94	3-3.33	3.27	0.94	2.98-3.55
I expected to get critical feedback from them	trainee	3.26	0.97	3.09-3.43	3.36	0.98	3.06-3.65

Overall the means are higher for the usefulness of text than numerical feedback (n=874, text mean=3.71, SD=0.94, CI=3.65-3.77; numerical mean=3.35, SD=0.96, CI=3.28-3.41; paired t-test $t(873)=10.88$, $p<0.001$, Cohen's $d=0.39$).

3.8 Questions of validity

Trainees and supervisors were asked if they felt feedback was trustworthy. Trainees were more positive ($F(1,269)=4.364$, $p<0.05$, $f=0.13$), but there was no difference between tools.

3.8.1 Basis of feedback

Pilot work had identified trainee concerns that feedback was not based on direct observation of their behaviour. In this study 31% of trainees agreed or strongly agreed that this was a concern, although the overall mean was just 2.82 (SD=1.20, CI=2.64-3.00). There was a difference between tools however ($F(1,169)=5.108$, $p<0.05$, $f=0.17$), with TAB trainees expressing more concern. However, raters using both tools felt that they had sufficient experience of working with trainees to give feedback (mean=3.79, SD=0.93, CI=3.71-3.86; 70% agreeing or strongly agreeing).

Table 5 gives rater responses to the question ‘What did you base your feedback on?’. The proportions of responses are comparable for both groups of raters. The most frequently reported basis of feedback was ‘direct observation of behaviour on several occasions’, although 3% of raters selected only ‘Direct observation on one occasion’, and 2% did not select either 'direct observation' option.

The second most frequent response was ‘discussion with colleagues’, and other indirect sources were also frequently indicated, such as an absence of hearing negative comments about a doctor, or inference from behaviour not directly referenced by the feedback tools. ‘Other’ sources included comments from patients, comments from nurses, formal educational contact, and simply ‘working with them’.

Table 5. Percentages of responses to 'What did you base your feedback on?'. Respondents could tick more than one item.

	Mini-PAT (n=472)	TAB (n=155)
Direct observation on several occasions	95	95
Discussion with colleagues	48	42
Inference from other observed behaviour	23	23
Direct observation on one occasion	17	8
(Direct observation on one occasion only)	(4)	(1)
Absence of negative reports	16	11
Personal ('off-duty') knowledge of the doctor	9	8
Other	4	2

3.8.2 Selection of raters

Concerns around potential bias in doctors' selection of raters had also been raised in pilot work. Specifically there were concerns that trainees would select raters for reasons other than their having good access to their practice. Trainees were therefore asked to indicate their agreement that different factors had influenced their selection of raters. The highest rated was 'I know they have experience of my work' (mean=4.50, SD=0.69, CI=4.40-4.61), followed by 'I get on with them as a person' (mean=3.70, SD=0.91, CI=3.57-3.84), 'I expect to get critical feedback' (mean=3.29, SD=0.98, CI=3.14-3.44) and 'I expect to get positive feedback' (mean=3.19, SD=0.93, CI=3.04-3.33). There were no differences between the tools.

There were also 'other' responses added by respondents in a free text area on the questionnaire. These included selecting raters expected to return the form (implying others would not), or working with a small team and so having no choice of raters. Other responses referred to selecting people whose knowledge and expertise they respected, ensuring coverage from different grades, having been supervised by a rater, and raters' honesty and punctuality.

4. Discussion

Opinions of the ease of use, usefulness and validity of two multi-source feedback tools – the mainly textual, relationship-oriented TAB, and the mainly numerical and more wide-ranging mini-PAT – were collected by postal questionnaire. Analysis compared attitudes towards the two tools, and of the three groups who used them (trainees, raters and supervisors).

The differences found are small, suggesting that attitudes to the two tools are generally similar. TAB was though felt to be more useful in the areas of communication and professionalism for which it was originally developed[5]. Its greater provision for text was also reflected in its users' finding textual feedback more useful than those of mini-PAT. Textual feedback was considered to be more useful by trainee and supervisor users of both tools, supporting findings in the non-healthcare MSF literature[24]. However, raters reported that scale feedback was more useful. This may indicate a general preference to *give* quick, 'broad brush' feedback, but to *receive* detailed, personally-tailored feedback.

Despite positive attitudes to MSF in principle, users of both tools had low expectations of its effectiveness, and nearly a third of trainees did not anticipate changing in response to feedback. The influence of a facilitator in the effectiveness of MSF has recently been established in qualitative research[25] and it may be that trainees who were not planning changes had supervisors who were less inclined to respond to feedback. Faculty development in the provision of feedback may therefore be key[26]. There are also findings in the literature that feedback effectiveness is affected by individual differences[11 27].

The usefulness of feedback will depend on the extent to which feedback is based on raters' direct observation and knowledge of that behaviour. While the perceived validity was the main reason for selecting raters, pragmatic considerations were also relevant, such as raters being available to complete the tool in time. Interpersonal relationships also played a part. While the literature suggests that selection of raters does not affect ratings[28 29], all raters are not equal.

Trainees indicated concerns that feedback may not be based on direct knowledge. While the vast majority of raters said they based feedback on direct observation, a large minority were also using indirect evidence. This leads to the concern that rather than assessing observed behaviour responses may at best reflect global views, at worst biased preconceptions. Data from interviews with raters in the USA reinforces this[30], illustrating that the evidence they use is not always behavioural, but often extrapolated from holistic judgements.

Feedback has been found to vary with factors other than the rated behaviour, including the perceived stakes of feedback[31] (meaning summative assessments may elicit more lenient responses than formative ones), the response format of a tool[32], and rater qualities such as their professional group and seniority[8] or mood[33]. Interpersonal relationships can affect not just the ratings given, but the degree to which feedback items are completed at all[34]. These effects all indicate that feedback generation is not just passively mapping to a scale or a text box, it is an active, cognitive process and as such may be open to other cognitive biases[35 36].

Concerns over the reliability and validity of MSF have been raised in the past[12]. There is an epistemological, not just a methodological question whether the breadth of behaviour addressed can be 'measured' in the way in which other constructs are.

Recipients' preference for textual feedback is important here. Scale-based scores can be arithmetically aggregated, but even if those scores stem from appropriate knowledge, the value for recipients is in the heterogeneity of voices heard in feedback.

MSF has value for a trainee's learning, but if MSF is to be used for career-defining assessments, it should take account of the cognitive and social-cognitive context in which it is generated[37]. There are alternative, qualitative approaches to multi-source feedback and appraisal which may capitalise on the strengths of MSF[38 39 40], which may provide the complexity of feedback which recipients value, while being methodologically more simple.

4.1 Limitations of the study

The study had some limitations. The practical difficulties of accessing the populations meant response rates were uncertain, although the raw 40-50% of questionnaires which were returned is a healthy minimum. When the difficulties of reaching the populations involved are considered, the effective rate is likely to be higher.

The questionnaire used was revised from pilot work, and while its content validity was established, the high inter-correlations between items may indicate that attitudes towards tools such as this are prone to a halo effect.

There is a risk of statistical error in the low power of the analysis, stemming in part from unequal sample sizes, in part from small effect sizes. The practical relevance of the small effect sizes found may be questioned, but as Cohen has stated, effect size is a matter of context[41] and while the differences here are not conclusive, they do appear to reflect the different approaches of the tools, and the different relationships the groups have to them.

Finally, the UK Foundation Programme is more established since the study, and the MSF tools have changed in their delivery, although only slightly in content (of mini-PAT). However, the findings as they relate to differences between the content and format of the tools remain relevant.

Concerns about raters' exposure to doctors' behaviour may be more salient in the UK today, as the European Working Time Directive limits the amount of time trainees spend on wards and hence the opportunities for staff members to witness trainees' practice.

5. Conclusion

The study has identified some key issues around user perceptions of multi-source feedback:

- Many users did not feel the feedback tools were useful for education and development, or the identification of doctors in difficulty.

- Raters may prefer to give quick, numerical feedback, but trainees and supervisors find detailed textual feedback more useful.

- Raters may give feedback based on indirect sources of information such as discussions with colleagues in addition to their own direct observation.

- High-stakes applications of MSF in assessment and revalidation should consider the effects of context, and the psychological processes involved in feedback generation to ensure fairness, validity and reliability.

References

[1] Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. *Medical Teacher* 2006; **28**(7): e185-e191

[2] Davies H, Archer J, Bateman A, Dewar S, Crossley J, Grant J, Southgate L. Specialty-specific multi-source feedback: assuring validity, informing training. *Medical Education* 2008; **42**: 1014-1020

[3] Carr S. The Foundation Programme assessment tools: An opportunity to enhance feedback to trainees? *Postgraduate Medical Journal* 2006; **82**:576-579

[4] Archer J, Norcini J, Southgate L, Heard S, Davies H. mini-PAT (Peer Assessment Tool): A valid component of a national assessment programme in the UK? *Advances in Health Sciences Education* 2008; **13**(2):181-92.

[5] Whitehouse A, Hassell A, Wood L, Wall D, Walzman M, Campbell I. Development and reliability testing of TAB a form for 360 degree assessment of Senior House Officers' professional behaviour, as specified by the General Medical Council. *Medical Teacher* 2005; **27**(3):252-8

[6] Wood L, Wall D, Bullock A, Hassell A, Whitehouse A, Campbell I. 'Team Observation': A Six Year Study of the development and use of multi-source feedback (360 degree assessment) in obstetrics and gynaecology training the United Kingdom. *Medical Teacher* 2006; 28 (7): e177-e184

[7] Archer J, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005; **330**: 1251-1253

[8] Bullock AD, Hassell A, Markham WA, Wall DW, Whitehouse AB. How ratings vary by staff group in a multi-source feedback assessment of junior doctors. *Medical Education* 2009; 43: 516-520

[9] Whitehouse A, Hassell A, Bullock A, Wood L, Wall D. 360 degree assessment (multisource feedback) of UK trainee doctors: Field testing of team assessment of behaviours (TAB) *Medical Teacher* 2007; 29: 171-176

[10] Sargeant JM, Mann KV, Ferrier SN. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness *Medical Education* 2005; **39**: 497-504

[11] Sargeant JM, Mann KV, Sinclair D, van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Advances in Health Sciences Education* 2008; **13**(3): 275-288

[12] Hesketh EA, Anderson F, Bagnall GM, Driver CP, Johnston DA, Marshall D, Needham G, Orr G, Walker K. Using a 360° diagnostic screening tool to provide an evidence trail of junior doctor performance throughout their first postgraduate year. *Medical Teacher* 2005; **27**(3): 219-233

[13] Rees C, Shepherd M. The acceptability of 360-degree judgements as a method of assessing undergraduate medical students' personal and professional behaviours. *Medical Education* 2005; **39**: 49-57.

[14] Murphy DJ, Bruce DA, Mercer SW, Eva KW. The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Advances in Health Science Education: Theory and Practice* 2008;**14**: 219-232

[15] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*1989; **13**(3), 319-340

[16] Wu J-H, Shen W-S, Lin L-M, Greenes RA, Bates DW. Testing the technology acceptance model for evaluating healthcare professionals' intention to use an adverse event reporting system. *International Journal for Quality in Health Care*. 2008; **20**(2):123-9

[17] Hrisos S, Illing J, Burford B. Portfolio learning for foundation doctors: early feedback on its use in the clinical workplace. *Medical Education* 2008; **42**(2): 214-223

[18] Burford B, Illing J, Hrisos S, Archer J, van Zwanenberg T, Livingston M. *An evaluation of a formative 360 degree feedback tool for trainee doctors: Preliminary results.* Presentation at ASME Annual Scientific Meeting, Newcastle, July 2005

[19] UK Departments of Health *The Foundation Learning Portfolio* (revised August 2009) [http://www.foundationprogramme.nhs.uk/download.asp?file=Learning_PortfolioV3-1.pdf Accessed 19 August 2009]

[20] Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. Needham Heights, MA: Allyn and Brown, 2001. Section 8.5.2.2, Unequal sample sizes, p.296-297

[21] Cohen J. *Statistical power analysis in the behavioural sciences.* Second edition. Hillsdale, NJ: Erlbaum, 1988. Section 8.2, The effect size index: f, p. 274-288

[22] Nakagawa S. A farewell to Bonferroni: the problems of low statistical power and publication bias *Behavioral Ecology* 2004 15 No. 6: 1044–1045

[23] General Medical Council. *Good Medical Practice*. (3rd edition) London: GMC, 2001

[24] Smither JW, Walker AG. Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time. *Journal of Applied Psychology* 2004; **89**: 575-581

[25] Overeem K, Wollersheim H, Driessen E, Lombarts K, van de Ven G, Grol R, Arah O. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. *Medical Education* 2009; **43**: 874-882

[26] Bing-You, RG and Trowbridge, RL. Why medical educators may be failing at feedback. *JAMA* 2009; **302**: 1330-1331

[27] Smither JW, London M, Reilly RR. Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology* 2005; **58**: 33-66

[28] Ramsey PG, Weinrich MJ, Carline JD, Inui TS, Larson EB, Logerfo JP. Use of peer ratings to evaluate physician performance. *JAMA* 1993; **269**: 1655-1660

[29] Lurie SJ, Nofziger AC, Meldrum S, Mooney, Epstein RM. Effects of rater selection on peer assessment among medical students. *Medical Education* 2006; **40**: 1099-1097

[30] Mazor KM, Canavan C, Farrell M, Margolis MJ, Clauser BE. Collecting validity evidence for an assessment of professionalism: Findings from think-aloud interviews. *Academic Medicine* 2008; **83**(10 suppl): s9-s12

[31] Harris MM, Smith DE, Champagne D. A field study of performance appraisal purpose: research- versus administrative-based ratings. *Personnel Psychology* 1995; **48**: 151-160

[32] Plous S. *The Psychology of Judgment and Decision Making*. New York : McGraw-Hill, 1993

[33] Antonioni D, Park H. The relationship between rater affect and three sources of 360-degree feedback ratings. *Journal of Management* 2001; **27**: 479-495

[34] Mazor K, Clauser BE, Holtman M, Margolis MJ. Evaluation of missing data in an assessment of professional behaviors. *Academic Medicine* 2007; **82** (10 suppl): s44-s47

[35] Gilovich T, Griffin D, Kahneman D. (Eds.). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press, 2002

[36] Eva K, Norman G. Heuristics and biases − a biased perspective on clinical reasoning *Medical Education* 2005; **39**: 870 - 872

[37] Feldman JM. On the synergy between theory and application: Social cognition and performance appraisal. In RS Wyer Jr and TK Srull (Eds.) *Handbook of social cognition* (2nd ed., Vol. 2) Hillsdale, NJ: Erlbaum. 1994

[38] Tintinalli JE. Evaluation of emergency medicine residents by nurses. *Academic Medicine* 1989 **64**: 49-50

[39] Garbett R, Hardy S, Manley K, Titchen A, McCormack B. Developing a qualitative approach to 360-degree feedback to aid understanding and development of clinical expertise. *Journal of Nursing Management* 2007; **15**: 342-347

[40] McLellan H, Bateman H, Bailey P. The place of 360 degree appraisal within a team approach to professional development. *Journal of Interprofessional Care* 2005; **19**: 137-148

[41] Cohen J. *Statistical power analysis in the behavioural sciences* Second edition. Hillsdale, NJ: Erlbaum, 1988. Section 11.2 Effect size, p.531-535