Developing a multisource feedback tool for postgraduate medical educational supervisors

Julian Archer MBBS MRCPCH MEd PhD FAcadMed

Clinical Senior Lecturer in Clinical Education at the

Peninsula College of Medicine and Dentistry, Plymouth, UK

Tim Swanwick MA (Cantab) MBBS FRCGP MA FAcadMed

Dean of Professional Development in the London Deanery, London, UK and Visiting
Professor of Medical Education at the University of Bedfordshire, UK

Daniel Smith MSc

Research and Analytics Manager, London Deanery, London, UK

Introduction

The effective supervision of doctors in training has been linked to good patient
outcomes. Supervisors play a key role in the development of doctors in clinical
training programmes both in the oversight of their day-to-day practice but also in the
support and orchestration of their learning experiences (Kilminster S, Cottrell D et al.
2007). In the UK there has been an attempt to make a clear distinction between
these two activities of 'clinical' and 'educational' supervision. Educational supervision
may take place in the context of a training placement or a training programme. In
both cases, the primacy of focus is on the educational trajectory of the trainee over a
defined period of time against objectives negotiated between trainer and trainee
within the context of institutional curricula that frame their relationship.

Over the last few years, regulatory requirements and policy initiatives have placed a
significant emphasis on the 'professionalisation' of medical educators, particularly in
the postgraduate arena, where there has been an inexorable growth in accountability
not only to patients and the service, but to the junior doctors themselves (Swanwick T
2008). Coupled with an international trend of widening scope and improved rigour of
relicensing and recertification processes, there is an increasing need for supervisors
to be able to demonstrate that they can fulfil the requirements of their role and in
doing so, to actively seek feedback on their performance.

In this paper we report on the development of a web-based multi-source feedback (MSF) tool for educational supervisors in the London Deanery[1] - an organisation responsible for 12,500 doctors in training programmes constituting around 25% of the UK's trainee population. The development of the MSF is part of a broader integrated faculty development strategy which includes a requirement for supervisors to undergo periodic appraisal for their educational role against a professional development framework (London Deanery 2009; Swanwick T, McKimm J et al. 2010). The use of professional and personal judgements to evaluate aspects of performance in the workplace is well established in industry, medicine (Norcini 2003) and medical education (Archer, Norcini et al. 2005). It has also been adopted as a tool for the evaluation of undergraduate medical faculty (Copeland HL and Hewson MG 2000). In the context of postgraduate medical education it was envisaged that a valid and reliable MSF tool would provide supervisors with a unique and useful means to evidence their educational competence. Elsewhere, in a companion paper, we report on the tool's first 12 months of use and a programme of improvement and development.

Methods

A multi-methodological approach to development was commissioned from researchers at the Peninsula College of Medicine and Dentistry. Central to the success of any assessment instrument is assuring its validity; a complex concept that Messick describes as an overall evaluative judgment of the degree to which empirical evidence supports the adequacy and appropriateness of interpretations and actions based on assessment scores (Messick 1989). Downing recommends the collection of validity evidence under five headings; content, response process, internal structure, relationship to other variables and consequences (Downing SM 2003). It is the first three of these sources of evidence that provide a framework for the methods and analysis adopted in this study.

---

[1] The rationale for restricting the target users of the instrument to educational supervisors relates to the fact that this job role can more easily be defined than that of 'clinical supervisor' which is often difficult to separate out from the 'background radiation' of clinical supervision provided by every clinician with whom a trainee comes into contact.

Content validity

Assuring the content validity of the instrument was achieved through two key methods. First, key literature sources were identified through standard literature databases and a content analysis undertaken. This brought together the profession's expectations of the role of an educational supervisor through the examination of key documents, (Department of Health 2007; Department of Health 2007; Postgraduate Medical Education and Training Board 2008) with a review of the literature on educational supervision of which Kilminster et al (Kilminster S, Cottrell D et al. 2007) and Glanz and Neville (Glanz J and RF. 1997) provide illustrative summaries both within and outside the medical context.

Second, focus groups of key stakeholders were held (with educational supervisors, supervisees, London Deanery and lay representatives), the purpose of which was to capture what was perceived to be important for a successful supervisory relationship. The structural framework of the focus groups was informed by the literature review in anticipation that the data obtained would take the evaluation instrument beyond a simple satisfaction questionnaire and capture potentially richer information about the professional relationship so providing the educational supervisor with useful feedback to inform their professional development.

The data from both the literature and the focus groups were analysed by two of the researchers using a process of interpretative thematic analysis. Later, free text comments from the pilot of the live instrument were incorporated and a combined thematic framework agreed

Analyses of the triangulated data from the first phase of the study led to the formulation of 21 items and a prototype instrument was tested for face validity with a group of volunteer educational supervisors Each item was placed against a six-point scale (1-6 with an unable to comment option that was not scored), developed from a previous review of the literature (Archer J 2007). The MSF was piloted online with a volunteer cohort of educational supervisors over a three month period inviting ratings by trainees and asking educational supervisors to rate themselves. The performance of all items was scrutinised and a review undertaken of the free text in order to help identify recurring themes that might not be specifically asked within the items and to support validation of existing items. Demographic data were collected about both the supervisor and the trainee in order to explore the possibility that scores were being moderated (Baron RM and Kenny DA 1986) by these variables.

Response process

A review of the delivery system was undertaken to evaluate the impact of the process on validity. This was aided by the fact that a standard online MSF product was used with a proven track record. Amendments to the online processes were implemented as an ongoing and integral part of the development phase. It was assumed that as part of the pilot self sign up by volunteers would be necessary. Consent was gained through the completion of an online section prior to continuing to the assessment forms. Educational supervisors were asked to complete a supervisee form containing all the contact details for current and recent (in the last 2 years) supervisees. They were also asked to complete a self assessment form and demographic data form. Supervisees were then contacted by email asking for them to complete an evaluation instrument form. The data from all sources were then prepared in an report, providing aggregated means across all supervisees' ratings for each item and for both scale scores, where the scale score was a mean of a group of items, together with the self assessment scores for the same items and scales. Free text comments were also fed back, anonymised but verbatim.

Internal structure

Psychometric theories offer a range of approaches that try to explain the behaviour of an instrument and provide evidence for validity. In this study, item analysis was undertaken to explore the appropriateness and behaviour of the items within the new instrument. Analyses included item-total and item-item correlations, factor analyses, and internal consistency and response rates. Sources of bias were also explored by regression analyses looking for characteristics that might systematically affect scores independently of the educational supervisors.

Reliability was specifically explored using Generalisability theory (Cronbach L and Shavelson RJ 2004). The naturalistic design of the study means that the model for the analysis was fundamentally nested; but with some crossing of assessors this allowed for the estimation of the variance for intra-rater reliability as part of the overall model. Nesting means that in the vast majority of cases assessors were unique to each educational supervisor. This naturalistic design does limit the variance components that can be calculated from the model but this was addressed in some respects by looking at regression analyses as discussed.

The 'G' study was conducted using the standard error of the measure to calculate 95% confidence intervals. The square root of the measurement error (when estimated for varying numbers of assessors and/or events/items) constitutes the standard error of measurement ($\sqrt{}$measurement error). 95% confidence intervals around the standard error are equal to the standard error of measurement multiplied by 1.96 and are added to and subtracted from an individual's mean rating. This method provides a measure of the precision of the score achieved on the scale. If educational supervisors are seen as doing well 95% CIs help to reduce the sampling needed to achieve a reliable result however for those nearer or below the criterion standard further sampling is ideally needed.

Results

One hundred and five of the 128 educational supervisors (82%) initially approached to participate were assessed by 634 trainees over a 3 month period. Overall 655 forms were completed, 21 trainees completing 2 forms (one on each of two different supervisors). On average each educational supervisors had feedback from 6 trainees (range 1-9). An aggregate score (mean) was calculated for each form and for each educational supervisor. Scores from trainees ranged from 1.4 to 6.0. Educational supervisors aggregate scores from their trainees ranged from 4.33 to 6.0 (mean 5.54, SD 2.69). Educational supervisors' self scores correlated poorly; with educational supervisors generally scoring themselves lower than their trainees ($R = 0.234$, $P = 0.019$). Both supervisors and trainees came from a range of organisations and from diverse ethnic backgrounds (see Table 1).

Table 1

Demography of pilot participants

Educational supervisor		
	Frequency	*%*
Ethnicity		
White	72	69
Black	14	13
Asian	3	3
Chinese	5	5
All other	6	6

Not stated	5	5
Total	105	100
Environment		
Acute	75	71
Mental health	23	22
Primary care	2	2
No data	5	5
Total	105	100
Trainee		
	Frequency	*%*
Ethnicity		
White	338	53
Black	182	29
Asian	31	5
Chinese	21	3
All other	26	4
Not stated	36	6
Total	634	100
Environment		
Acute	480	76
Mental health	136	22
Primary care	18	3
Total	634	100
Stage of training		
Foundation	115	18
Specialty	519	82
Total	634	100

Content validity

Data from the literature review and focus groups were analysed using an interpretative thematic approach which resulted in 21 items. These items are listed in Tables 2 & 3. The items were then piloted against the 6 point scale using the developed online system.

On analysis of the pilot data, overall the instrument was found to be suitable for factor analysis (KMO = 0.928; Bartlett test significant, $p < 0.001$) with a two factor solution accounting for 76.5% of the variance. These factors were provisionally entitled 'the process of educational supervision' and 'going the extra mile'. See Table 2. With a

further analysis of the free text comments from the pilot phase of the study the two factors were further developed to include three sub-themes each: Within 'the process of educational supervision'; 'knowledge of the process', 'attitude to supervision', 'taking an interest in me professionally', and under 'going the extra mile'; 'working outside the box', 'making time', and 'acting as a role model'.

Table 2

Principle components factor analysis

Item		Component	
		1	2
1	Ability to remain up to date about your training scheme	.410	.571
2	Genuine interest in your portfolio	.500	.627
3	Approachability	.281	.830
4	Enthusiasm	.594	.612
5	Ability to inspire you	.757	.458
6	Ability to seek help form other sources	.614	.368
7	Ability to challenge you	.917	.023
8	Willingness to act to resolve problems in a timely manner	.569	.577
9	Ability to give constructive feedback	.745	.331
10	Communication skills	.532	.528
11	Ability to communicate with your clinical supervisors	.459	.557
12	Encouragement towards you achieving excellence	.718	.467
13	Ability to take your supervision beyond a tick box exercise	.723	.463
14	Honest and integrity	.420	.693
15	Ability to assure privacy and where appropriate, confidentiality	.519	.525
16	Ability to make time for you	.146	.880
17	Commitment to rearrange meetings they have cancelled	.138	.834
18	Interest in you as an individual	.546	.641
19	Ability to be your advocate	.451	.598
20	Ability to offer practical tailored advice for your longer term career planning	.813	.284
21	Overall how do you rate your educational supervisor	.692	531

Note: final scales based on rotated components with the three excluded items removed

Inter-item correlations varied from 0.195 to 0.862 (p<0.001). Item-total correlations were all above 0.647 (p<0.001). All items were answered 85% of the time, but three questions were answered significantly less frequently than the other items. How is your educational supervisor in their ability to (i) *seek help from other sources*, (ii)

ability to communicate with your clinical supervisors, (iii) *commitment to re-arrange meetings they have cancelled*? See Table 3.

Table 3

Non-response rates

Item		%
1	Ability to remain up to date about your training scheme	0.2
2	Genuine interest in your portfolio	1.7
3	Approachability	0.2
4	Enthusiasm	0.2
5	Ability to inspire you	0.6
6	Ability to seek help form other sources	12.1
7	Ability to challenge you	2.6
8	Willingness to act to resolve problems in a timely manner	3.7
9	Ability to give constructive feedback	0.2
10	Communication skills	0.2
11	Ability to communicate with your clinical supervisors	13.3
12	Encouragement towards you achieving excellence	0.5
13	Ability to take your supervision beyond a tick box exercise	1.2
14	Honest and integrity	0.5
15	Ability to assure privacy and where appropriate, confidentiality	2.6
16	Ability to make time for you	0.2
17	Commitment to rearrange meetings they have cancelled	13.2
18	Interest in you as an individual	0.5
19	Ability to be your advocate	5.5
20	Ability to offer practical tailored advice for your longer term career planning	2.9
21	Overall how do you rate your educational supervisor	0

The naturalistic design of the study allowed for the calculation of variance estimates attributable to the educational supervisor (true difference between educational supervisors), the trainee as the assessor (intra-rater variance) and error (all other un-attributable variance). Using these variance components (see table 4) it is possible to calculate reliability with increasing numbers of assessors. This is summarised in table 5 showing both Phi or D coefficients and using the error components - calculated 95% confidence intervals (95%CI).

Table 4

Variance estimates

Component	Variance estimate	%
Var (Educational Supervisor)	.025	9
Var (Trainee/Assessor)	.043	16
Var (Error)	.206	75

Reliability was lower than might be expected with such an instrument. Generalisability 95% CIs allow 'precision' to be placed around scores in relation to the number of assessors contributing to it. The practical implication of this use of G theory is demonstrated in figure 1. Generalisability analysis data, summarised in Table 5 suggests that using 95% CIs, as few as 3 trainees feeding back to their educational supervisors would allow decisions to be made about whether the educational supervisor's performance was satisfactory or not. This was the case for all but one supervisor in this cohort. For this one supervisor further sampling would be required.

Table 5

D study with 95% confidence intervals

Number of assessors	Phi/D	95% CI
1	0.1	1.0
2	0.2	0.7
3	0.2	0.6
4	0.3	0.5
5	0.3	0.4
6	0.4	0.4
7	0.4	0.4
8	0.4	0.3
9	0.5	0.3
10	0.5	0.3
11	0.5	0.3
12	0.5	0.3
13	0.6	0.3
14	0.6	0.3
15	0.6	0.3
16	0.6	0.2
17	0.6	0.2

18	0.6	0.2
19	0.7	0.2
20	0.7	0.2

The only possible systematic influence on scores was the ethnicity of the educational supervisor, White and Asian supervisors scoring lower than Black, Chinese and 'Other' ethnic groups (White t=-2.7, p<0.05, Asian t=-5.1, p<0.001). However this was small and when seen in a study with the pilot study did not capture place of primary medical qualification, which is confounded with ethnicity.

Discussion

This paper describes the development of an instrument that will enable educational supervisors to receive valid and reliable feedback from trainees on the supervision received.

While the study described has many strengthens, including a multi-methodological approach combining quantitative with qualitative data, there are limitations. The literature was not systematically reviewed and analysed. This was a pragmatic decision and unlikely to have affected the results from the review. The decision was made as there were recent and significant systematic reviews already in existence. Our narrative review was to inform the focus groups more than the study overall and as expected this subsequent methodology yielded similar yet richer results. The instrument closely reflected the thematic outcomes of the focus groups and bias was limited by triangulation of results and two researchers being involved. However the question framework was an interpretative approach and was therefore open to researcher bias. The group were volunteers so the study was subject to selection bias in relation, for instance, to engagement with the role and specialty. In further studies the reliability and systematic biases as well as the ability for the instrument to highlight excellent and poor performance will need to be re-examined.

As a result of the development process described above, a number of modifications were made to the final MSF instrument before it went live online in July 2010.

The three items that had been answered significantly less frequently have subsequently been removed from the questionnaire resulting in a final 18 item

instrument. The first two of these ask trainees to rate their educational supervisors in their ability to undertake tasks that would not necessarily be apparent to the trainee. They are secondary events and as such are not suitable for Multisource Feedback (MSF) technology. The third question was not clearly as relevant to this cohort but may be of significance subsequently.

In the rollout of the instrument, a requirement for five respondents was introduced in order to generate a report. This was a pragmatic decision as although generalisability theory analysis showed that the instrument was reliable with lower numbers of - as few as 3 trainees provided a 95% CI of 0.6 - it was felt that a greater number of trainees responding would strengthen validity overall, have the potential to increase the number of free text comments provided and further protect anonymity.

Minor modifications were also made to the presentation of the online framework including presenting the questions under the two factors identified labelled as 'personal attributes' and 'challenge and support'. Branding was applied and clear instructions, sample reports and FAQs provided. Full details about the final instrument and a sample report can be found at:
http://www.faculty.londondeanery.ac.uk/supervisor-MSF

Conclusion

This paper summarises the development of a novel instrument allowing trainees to feed back to their educational supervisors about their experience of supervision. The study assures content validity through a robust developmental framework for the items in the instrument and uses psychometric techniques that further support validity including reliability at feasible levels for implementation. The tool has great potential allowing supervisors to evidence their ability. This is particularly timely in the UK with the development of a regulatory requirement for the accreditation of doctors with a formal responsibility for teaching or training in the clinical environment (General Medical Council 2012)

A programme of further development has recently been completed including a review of content validity through the analysis of free text responses obtained in the live environment, a repeat of item analyses with a larger and more heterogeneous group or participants, and consequential validity explored through a qualitative study of user

feedback. A summary of these analyses, data from the instrument's first full year of use and a summary of further developments are presented in a companion paper.

Acknowledgements

Ethical approval

Ethical approval was sought from the Chair of the King's College Hospital Research Ethics Committee, London. Advice was received stating that the study did not need formal approval as it was considered to be a 'service evaluation'.

References

Archer J. Multisource Feedback to Assess Doctor's Performance in the Workplace. University of Sheffield, Sheffield, 2007.

Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training 10.1136/bmj.38447.610451.8F. *BMJ.* 2005;**330**:1251-1253.

Baron RM, Kenny DA. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*. 1986;**51**:1173-1182.

Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical centre. *Academic Medicine*. 2000;**75**:161-166.

Cronbach L, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. *Educ Psychol Measurem*. 2004;**64**:391-418.

Department of Health. A Guide to Postgraduate Medical Education in the UK (The Gold Guide). MMC, 2007.

Department of Health. Modernising Medical Careers Web Site. *www.mmc.org.uk*, 2007.

Downing SM. Validity: on the meaningful interpretation of assessment data. *Medical Education*. 2003;**37**:830-837.

General Medical Council. Recognition and approval of trainers: a consultation. 2012:Accessed 27th January 2012.

Glanz J, RF. N. *Educational supervision: perspectives, Issues and Controvedrsies*. Norwood, MA Christopher-Gordon Publishers Inc; 1997.

London Deanery. *Professional Development Framework for Supervisors*. http://www.faculty.londondeanery.ac.uk/professional-development-framework-for-supervisors Accessed 18th November 2011; 2009

Kilminster S, Cottrell D, Grant J, B. J. AMEE Guide No. 27: Effective educational and clinical supervision. *Medical Teacher*. 2007;**29**:2 - 19.

Messick S. Validity. In: Linn R, ed. *Educational Measurement (3rd edn), American Council on Education*. Macmillan, Washington, 1989; 13-104.

Norcini J. Peer assessment of competence. *Med Educ*. 2003;**37**:539-543.

Postgraduate Medical Education and Training Board. *Generic Standards for Training*. London Postgraduate Medical Education and Training Board; 2008.

Swanwick T. See one, do one, then what? Faculty development in postgraduate medical education *Postgraduate Medical Journal* 2008;**84**:339-343.

Swanwick T, McKimm J, R. C. Introducing a professional development framework for postgraduate medical supervisors in secondary care: considerations, constraints and challenges *Postgraduate Medical Journal Postgraduate Medical Journal*. 2010;**86**:203-207.