Problems with using a supervisor's report as a form of summative assessment

Tim J Wilkinson, Winnie B Wade

The place of a supervisor report when used as a summative assessment of clinical workplace based learning is discussed

Within clinical medicine, the apprenticeship model is traditional, and highly valued. It relies on a close relationship between a supervisor and a trainee. When it comes to assessing the trainee, who better to ask than the supervisor? On the face of it, this approach makes good sense and has contributed to formalising ways of seeking such an opinion. As one source of feedback, such an opinion is highly valuable. In recent times, though, such reports are now being increasingly used as a form of summative assessment—that is, the basis on which decisions about the trainee's progress are made. This practice relies on the assumption that such a report is always a valid and reliable assessment method. We wish to challenge this assumption. This paper aims to distil and explain the fundamental flaws of this type of assessment, and offers an alternative solution that not only aids learning, but does so on the basis of more objective and unbiased information. We suggest that the use of a supervisor report change from that of an assessment *tool* to one that becomes a *summary* of the results of a variety of assessments.

WHAT IS A SUPERVISOR'S REPORT?

Typically, this is a form with a number of criteria deemed to be important for a trainee to achieve. The supervisor is asked to tick a box that best applies to that trainee's level of competence or achievement. For example, "the trainee has developed a level of knowledge commensurate with his or her level of training" or "the trainee is reliable" or "the trainee communicates well with patients and peers". Alongside these statements are levels of accomplishment, such as "below expectations, marginally below expectations, marginally above expectations, above expectations, well above expectations". There are many variations in the wording on these forms.

CURRENT USE

If such a report were working well as a form of assessment, we might expect it would be easy for a supervisor to report on

unsatisfactory or marginal attributes. We might also expect that trainees would not challenge the evidence underpinning such ratings and we might see some convergence in how these reports are structured. We, and others,^{1,2} have observed that these phenomena are not occurring.

Instead, we see that most trainees are rated as "above average", that rarely is anyone rated as unsatisfactory, and some trainees challenge comments that are less than favourable. Such challenges question the evidence on which the supervisor judgements are made.^{3,4}

The other phenomenon that we have observed is that there is a tendency to blame the supervisor for all of this—"if only they'd use the full range of the scale" or "Why don't the supervisors report on these areas of difficulty?" The frequent consequence is that the designers subject the form to further changes. A common response to this has been expansion of the number of categories—for example, by adding "borderline" or "marginal" or "barely meets expectations". Sometimes, as a consequence of further frustration, "borderline fail" and "borderline pass" might be added. When this does not work, the number of criteria and categories that are to be judged is expanded. With each step, the developer of the report form believes that he or she has solved the problem. Yet, it is our observation that each step along this path still may add to the frustration. On the other hand, for a trainee who is perceived to be performing well, such a form can be regarded as a satisfying experience for both supervisor and trainee, because all the feedback is positive, both parties are comfortable they are doing their jobs well, and the relationship remains collegial. The difficulties encountered in using these forms for trainees who are not performing well are the result of fundamental underlying problems with the attempt to use such information in this way.

PROBLEMS

We propose that the problems can be categorised into four main areas: conflict

of roles; assessor specificity; the halo effect; and assuming that objectivity arises from increasing the complexity of the report form.

Conflict of roles

When a supervisor is asked to be the only person to complete a supervisor's report assessment, this can create a conflict between the supervisor's role of facilitator of learning, and the role of judge of the effectiveness of that learning. Three roles are operating: the role of the trainee, the role of the supervisor as guide to learning, and the role of supervisor as judge. If all is going well, there may not be a conflict as the trainee is satisfied, and both roles of the supervisor are satisfied. All three roles are aligned. However, if the trainee is experiencing problems, these three viewpoints can come into conflict. This means that the "friendly" role of facilitator of learning conflicts with the "judging" role of assessor, and with the "learning" role of the trainee. This conflict not only affects assessment, it can also inhibit trainee learning. A supervisor can best help learning if the trainee can freely acknowledge weaknesses or gaps in learning. However, if a supervisor is also judging the trainee, then that trainee may wish to conceal such weaknesses. The task of helping a trainee find weaknesses in their abilities within a trusting relationship is vital to helping learning and central to the apprenticeship model.

This conflict of roles can also cause problems in trying to decide if difficulties are due to lack of learning or to lack of supervision. If learning has not occurred, there may be anxiety that there was inadequate supervision. This reason is frequently offered as an excuse by a failing trainee—"I wasn't given enough opportunities to learn". For the supervisor, "If I say he was unsatisfactory will he then challenge my ability to supervise?"

While skilled supervisors can often manage these conflicts, it does require a degree of experience and skill. Such skills in a supervisor are not automatically acquired. In such cases, there can be a real temptation to tick the satisfactory box and leave the problem to the next supervisor.

Assessor specificity

Within assessment, case specificity is a well-known phenomenon. An example can occur within assessments of clinical skills where performance with one patient may reflect particular aspects of the patient (or case) more than the competence of the trainee. We normally get around this problem by ensuring that clinical assessments have a variety of

patients and a variety of tasks. The aggregate performance across a range of patients gives a better picture of a trainee's ability. A similar phenomenon can occur with assessors. If the number of assessors is small, as is often the case,⁵ there is a risk that performance, as judged by one assessor, may tell more about the assessor than the trainee. When a trainee challenges a poor rating, the supervisor can be left in the uncomfortable position of having to defend his or her opinion against the trainee's opinion. This can create legal uncertainties and anxieties.⁶ If a supervisor's report is the opinion of one person, unrecognised biases may confound the assessment. We all have our idiosyncrasies—some supervisors look for particular attributes more than others. Sometimes what is being looked for may not be made overt. The way to reduce this risk is to aggregate the opinions of a range of assessors.

Halo effect

This refers to the tendency to rate someone highly on particular attributes when in fact they have strengths in other attributes.⁷ For example, if a trainee is friendly and gets on well with staff and patients, there can be a tendency to think they have good clinical examination skills or have a good knowledge base.⁸ Similarly, if a trainee is often late, there can be a tendency to generalise this negative perception into other aspects of his or her work. This is a well-described phenomenon, not just within medicine. People rarely seem to think of each other in mixed terms; instead we seem to see them as generally good or generally less good, across all categories of measurement. If there is only one rater, there is a greater risk of only a limited number of attributes being observed,⁵ and for the halo effect to occur, unbalanced and undetected.

Assuming that objectivity arises from increasing the complexity of the report form

Having clearly stated expectations and objectives is very helpful for learning, and is to be encouraged. It is tempting to conclude that translating these expectations down to the last detail on a form would lead to greater objectivity and thereby greater reliability in the assessment of these expectations. Furthermore, if there are problems arising from the subjectivity of the supervisor, or in the ability of such forms to detect unsatisfactory performance, then it is tempting to consider that altering the form may help. These observations are likely to have been drivers in making the report forms increasingly complex. There are two

reasons why this may not have solved the problems. The first relates to all the problems attributable to single assessors, such as conflict of roles, assessor specificity and the halo effect, as discussed above. The second problem is the observation from other forms of assessment whereby intricate checklists do not always enhance reliability. More global judgments can be just as reliable, provided such judgements are aggregated alongside judgements from a range of assessors. The phenomenon of making checklists too complicated in the hope that they might improve reliability has been described as objectification and contrasts with objectivity.⁹ The tendency to increase the number of categories or boxes to tick in supervisor's reports runs the risk of trivialisation and is a likely explanation for some checklist's inability to capture the "big picture", or the essence of performance expertise.^{10 11} It annoys the raters, undervalues their opinions, trivialises the tasks expected of trainees, and most importantly, does not solve any of the problems identified above.

SOLUTIONS

Although we have been critical of aspects of a supervisor's report, there are elements within the preceding process that are of great benefit and should be retained. These are: the emphasis placed on the personal relationship between a supervisor and trainee; the importance of regular meetings between supervisor and trainee; discussions between the supervisor and trainee about progress; the value of learning on the job; the importance of direct feedback from someone with more experience; the assistance a supervisor can provide in helping a trainee identify their gaps in learning; and the help a supervisor can provide in helping a trainee fill these gaps. All these aspects should be retained.

Within the form itself, there are often statements on the general areas that are of interest, and statements about the expected levels of achievement. These statements should be justified and evidence should be used to inform the statements. The report itself can serve as a very useful summary of the progress a trainee has made and can inform discussions when a trainee moves to a new supervisor. These are important and a useful guide to a trainee's learning.

The problems that need to be solved relate to the unreliability of undue reliance on information from just one assessor, and to the conflicts of roles. Principles of good assessment include the desirability of using a variety of assessment tools, a variety of assessors and a means to synthesise these pieces of information before making summative decisions.¹² The fundamental change that needs to occur to a supervisor report therefore is for its use to change from that of an assessment *tool* to one that becomes a *summary* of the results of a variety of assessments.

What types of assessment tools should contribute to this summary? Ideally, these are ones that are closely aligned to the areas of interest—that is, performance in the workplace. Fortunately there has been useful work in workplace based assessment, and development of a variety of tools.¹³ Examples include information derived from 10–20 colleagues' returns on a multi-source feedback (MSF) exercise,¹⁴ results from a series of mini clinical evaluation exercise (mini-CEX) assessments¹⁵ and other more novel assessment tools such as case-based discussions¹³ or direct observation of procedural skills.¹³ These tools are already in use in many countries worldwide and have four particular advantages: (1) they form very useful triggers for discussion between supervisors and trainees; (2) feasibility can be enhanced as they can be embedded within everyday work; (3) because they closely

Main points

- Supervisor's reports are being increasingly used as a form of summative assessment. This assumes that such a report is always a valid and reliable assessment method of assessment. We wish to challenge this assumption
- We propose that the fundamental flaws can be categorised into four main areas: conflict of roles, assessor specificity, the halo effect, and assuming that objectivity arises from increasing the complexity of the report form
- We suggest that the use of a supervisor report change from that of an assessment *tool* to one that becomes a *summary* of the results of a variety of assessments
- This retains the important collegial apprenticeship relationship but informs a judgement by drawing on a variety of sources of data from a variety of assessors

replicate every day work, they achieve greater validity; (4) regular use of direct observation and collection of data on performance contributes to a culture of quality improvement.¹⁶

This may mean each report should have a section that outlines the sources of evidence on which the summary was based, such as MSF, mini-CEX, reports from others, etc. If the only source of information were personal observation, then issues of unreliability of the judgment may become more obvious.

A further solution might be to consider separating the role of clinical supervisor from that of educational supervisor. A trainee would then have two supervisors: one for clinical matters who would probably be a senior doctor overseeing the care of the trainee's patients; and one for educational matters. The educational supervisor would be responsible for reviewing training and progress, and would also collate the results of other assessments—including information from the clinical supervisor. Needless to say, any supervisory role requires time and training—aspects that are not always plentiful in busy health services.

The most important benefit of these proposals is that it moves the dynamic between trainee and supervisor from a "one-versus-one" adversarial relationship back to a "one-beside-one" collegial relationship. For example, if the results of a multi-source feedback exercise suggest problems with inter-collegial relationships, then the supervisor can sit down with the trainee and discuss ways that these can be addressed, agree a plan of action and agree a date for additional data collection and review. The trainee-supervisor interaction would be further enhanced if the trainee undertook a self-assessment of the same criteria—areas of discrepancy between self- and external feedback are powerful triggers for meaningful discussions.

Some supervisors may already collect such data in an informal way. They may ask their colleagues for an opinion on the trainee's behaviour and performance and seek advice from allied health staff. While this is useful, it can lack the defensibility that is needed when difficult decisions about progress are to be made. The supervisor could still be accused of bias in selecting these opinions.

Until this happens, the supervisor report is doomed to continue to fail to achieve its intended purpose—and we predict that there will continue to be modifications of forms to alter criteria, to include or exclude more categories. These alterations will have no impact on a trainee who is performing well, but will not be robust enough to withstand challenge and scrutiny by those who perform less well. The accusations made of supervisors—"if only they would tick the Fail box more often"—will continue. The forms and reports will therefore continue to frustrate and the blame will continue to be laid at the supervisor's feet.

ACKNOWLEDGEMENTS

We are grateful to Professor Rufus Clarke for his helpful comments on an earlier draft

Postgrad Med J 2007;**83**:504–506.

doi: 10.1136/pgmj.2007.058982

Authors' affiliations

Tim J Wilkinson, Christchurch School of Medicine and Health Sciences, University of Otago, and Royal Australasian College of Physicians, New Zealand

Winnie B Wade, Royal College of Physicians, London, UK

Correspondence to: Professor Tim J Wilkinson, Christchurch School of Medicine and Health Sciences, C/- The Princess Margaret Hospital, PO Box 800, Christchurch, New Zealand; tim.wilkinson@chmeds.ac.nz

Funding source: Nil

Competing interest: Nil

Ethical approval: Not required

Contributions by authors: Both authors conceived and wrote the paper

REFERENCES

- 1 **Grant J**, Kilminster S, Jolly B, *et al*. Clinical supervision of SpRs: where does it happen, when does it happen and is it effective? *Med Educ* 2003;**37**:140–8.
- 2 **Holmboe ES**, Bowen JL, Green M, *et al*. Reforming internal medicine residency training. A report from the Society of General Internal Medicine's Task Force for Residency Reform. *J Gen Intern Med* 2005;**20**:1165–72.
- 3 **Kachalia A**, Studdert DM. Professional liability issues in graduate medical education. *JAMA* 2004;**292**:1051–6.
- 4 **Capozzi JD**, Rhodes R. Decisions regarding resident advancement and dismissal. *J Bone Joint Surg* 2005;**87**:2353–5.
- 5 **Daelmans HE**, Hoogenboom RJ, Donker AJ, *et al*. Effectiveness of clinical rotations as a learning environment for achieving competences. *Medical Teacher* 2004;**26**:305–12.
- 6 **Irby DM**, Milam S. The legal context for evaluating and dismissing medical students and residents. *Acad Med* 1989;**64**:639–43.
- 7 **Beckwith NE**, Lehmann DR. The importance of halo effects in multi-attribute attitude models. *J Marketing Res* 1975;**12**:265–75.
- 8 **Feeley TH**. Evidence of halo effects in student evaluations of communication instruction. *Communication Education* 2002;**51**:225–36.
- 9 **van der Vleuten CPM**, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* 1991;**25**:110–18.
- 10 **Wilkinson TJ**, Frampton CM, Thompson-Fawcett MW, *et al*. Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. *Acad Med* 2003;**78**:219–23.
- 11 **Hodges B**, Regheir G, McNaughton N, *et al*. OSCE checklists do not capture increasing levels of expertise. *Acad Med* 1999;**74**:1129–34.
- 12 **Epstein RM**. Assessment in Medical Education. *N Engl J Med* 2007;**356**:387–96.
- 13 **Modernising medical careers**. Assessment. <http://www.mmc.nhs.uk/pages/assessment> (Accessed 13 June 2007).
- 14 **Lockyer J**. Multisource feedback in the assessment of physician competencies. *Journal of Continuing Education in the Health Professions* 2003;**23**:4–12.
- 15 **Norcini JJ**, Blank LL, Duffy FD, *et al*. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med* 2003;**138**:476–81.
- 16 **Bolsin S**, Patrick A, Creati B, *et al*. Electronic incident reporting and professional monitoring transforms culture. *BMJ* 2004;**329**:51–2.