Original article

Effect of rating scales on scores given to junior doctors in multi-source feedback

Andrew Hassell,1,4 Alison Bullock,1,2 Andrew Whitehouse,1 Lawrence Wood,1,3 Peter Jones,1,4 David Wall

ABSTRACT

Background Multi-source feedback (MSF) has an established role in the workplace based assessment of doctors in training. Different models of MSF are currently used in different training programmes and settings. One important way in which these models differ is the rating scale on which assessors score the trainee. The aim of this study was to explore the effect of rating scale on MSF scores.

Methods Foundation Year 2 trainees in hospitals in the West Midlands underwent MSF using the validated MSF tool, team assessment of behaviour (TAB) in autumn 2005. Trainees were scored with TAB using one of four different rating scales, ranging from 3- to 9-point scales. Each participating hospital used only one rating scale. The proportions of trainees scored as having potential problems were related to the different rating scale used. Similarly, the proportions scored as ‘above expectations’ were compared. Assessors also completed a short questionnaire regarding the assessment.

Results 245 trainees underwent 2594 assessments. Longer rating scales were associated with a lower proportion of trainees awarded ‘problem’ scores and higher proportions of trainees scored as ‘above expectations’. Assessors generally reported no difficulties whichever rating scale they had used.

Conclusion Careful consideration, recognising its potential impact on assessment score, should be given to the rating scale used when instituting MSF within a system of workplace based assessment.

INTRODUCTION

Multi-source feedback (MSF) now has an established role in the assessment of postgraduate doctors in training in the UK and is utilised extensively in medical training and practice in many countries and settings.1–5 Models of MSF generally involve a number of colleagues (typically 8–12) completing a written form on which they score the individual in terms of a given set of behaviours. These assessments are then fed back to the individual, usually by a mentor or educational supervisor. MSF models differ in a number of ways, including the number and nature of domains in which the individual is scored and the rating scale and associated descriptors on which the score is made. In terms of the latter, rating scales which vary from 5-point to 9-point, with relevant descriptors, are currently in use in the UK.6 7

There is much in the market research and psychometric measurement literature which discusses rating scales, the number of points and the descriptors/anchors (eg, Preston and Colman8). By contrast, there is comparatively little within the medical education literature.9 14 Studies which have been published have tended to focus on score reliability rather than on the score awarded or the usefulness of the score.11 12

What is known is that all measurement has errors and that the error is partly attributable to the quality of the tool. Response scales and their descriptors and anchors are designed to help raters qualify their answers accurately and consistently. In a study intended to improve the design of rating scales, Davies found that “raters did not seem to know where a response of ‘not quite satisfied’ or ‘just barely satisfied’ should be placed”.15 The difference between these two points can have important consequences for medical trainees. One of the other challenges is in describing each point on the scale: for example, can we be sure that raters share an understanding of what ‘above expectations’ means? Is it clear whom expectations and based on what? In a comparison of school assessments across states in the USA, Perie found that differences in the percentage of students scoring at or above ‘proficient’ were partly due to the definition of ‘proficient’.14

In this study we have empirically investigated the impact of rating scales on the performance assessment of junior doctors (Foundation Year 2) using an MSF model—team assessment of behaviour (TAB), which has been previously found to show evidence of validity.15 Our specific aims were to investigate: (1) whether the length of the rating scale has any impact on the number of trainees scored as being of potential cause for concern; (2) whether the length of the rating scale has any impact on the number of trainees scored as above expectations; and (3) the views of assessors regarding the fitness for purpose of the different rating scales. The scales we investigated are those which are now used in several MSF tools in the UK.

METHODS

Team assessment of behaviour (TAB)

In autumn 2005, the West Midlands Deanery decided to pilot the use of TAB with Foundation Year 2 (FY2) trainees. FY2 doctors are in their first year after registration. The TAB process and scoring form have been presented elsewhere.6 15 The standard form can be accessed via http://www.foundationprogramme.nhs.uk/index.asp?page=home/ep-portfolio.15 For this study, the standard process and form was followed with one adjustment, the rating scale. Versions of the TAB form were produced with a 5-point rating scale (the published form, TAB5), a 4-point rating scale (TAB4), a 6-point rating scale

Postgraduate medical training in England

- In England, responsibility for the management of postgraduate medical training belongs to one of 13 ‘Deaneries’, each covering a geographical area of the country. The West Midlands Deanery covers the West Midlands, including Birmingham, Coventry and Stoke-on-Trent.
- All UK medical graduates complete a 2-year foundation programme before entering specialty training programmes. Typically, foundation programmes entail a series of six 4-month attachments in a variety of hospital settings. Commonly, one 4-month attachment is spent in general practice.
- Trainees successfully completing the first Foundation Year (FY1) are eligible for registration with the General Medical Council.

---

(TAB6, based on the scale used in the Mini-PAT (peer assessment tool) MSF system) and a 9-point rating scale (TAB9), based on the scale used in the MSF system developed and recommended by the Royal College of Physicians. Descriptors for each version of the scale were identical with those used in the original MSF tool from which the scale was derived. Our aim was to keep the size of all forms limited to the equivalent of one side of A4 paper. Consequently, whereas TAB5 and TAB4 had space for comments adjacent to each domain, TAB6 and TAB9 had only a single space for overall comments. On all forms, assessors rated FY2 doctors on four domains, as used in the classical version of TAB: maintaining trust/professional relationships with patients, verbal communication skills, team working and accessibility.

Participants

The FY2 trainees were based at one of 13 NHS Trust hospitals in the West Midlands. The 13 participating hospitals were allocated to one of four groups. The groupings were approximately matched for teaching/non-teaching hospitals, catchment area served by the hospital and number of trainees. Hospitals within each group were then allocated one version of TAB to pilot with the FY2 trainees.

Process

The usual process was then followed: trainees handed out at least 10 forms to colleagues (expected to include at least five nurses and the supervising consultant) who were asked to complete the form and return it in a sealed envelope to the nominated educational supervisor for the given trainee. The educational supervisor then summarised the results and met with the trainee to provide feedback. All completed forms were returned to the Deanery.

Questionnaires

Participating assessors were asked to complete a short questionnaire on the back of the TAB form. The questionnaire included a series of statements regarding the TAB process whereby respondents were asked to score their level of agreement on a 6-point scale (strongly disagree to strongly agree). Examples of these statements include: “It’s difficult to tell the difference between some points on the ratings scale” and “This form enabled me to congratulate good behaviour, where appropriate”. Other questions asked respondents to score the form according to defined purposes, on a 6-point scale. Examples include: “How effective is the form at providing worthwhile feedback?” and “How effective is the form in identifying trainees who may not be reaching the required professional standard?”

Analysis

Anonymised data from the forms were then entered onto NCSS 2004 (NCSS, Kaysville, Utah, USA) for analysis. To test the strength and significance of an inverse relationship between the proportions of concerns expressed in each domain and overall ratings, with the number of items on the TAB scale (three, four, six and nine items), a hierarchical loglinear model was used. In this model three factors (TAB rating scale version, domain and rating, that is, a ‘problem’ score (‘some or major concern’, ‘borderline or below expectations’ or ‘unsatisfactory’) or not, were fitted to the data and a stepdown technique was used to choose the best model. Hierarchical models are such that, if an interaction term is included in a model, then all other interaction terms made up of subsets of it are also included. For example, given three factors A, B, C, if the model contains the three-way interaction ABC, then all the two-way interactions (AB, BC, AC) and the main effects (A, B, C) will be included.

Statistical analysis was also used to analyse assessor views regarding the different rating scales. Responses of assessors using one rating scale were compared with those of assessors using another rating scale by means of the Kruskal–Wallis test.

RESULTS

The response rates are presented in table 1, the proportion of trainees scored as being ‘of concern’ according to the rating scale used are shown in table 2 and the proportion of trainees scored as ‘above expectations’ according to the rating scale used are given in table 3.

Response rates

Table 1 shows details of the number of trainees for whom data were available, the number of assessments and the number of questionnaire responses provided by assessors. It can be seen that there were 54–68 trainees for each version of TAB, representing a response rate of 63%–77%. The mean number of completed assessments per trainee was 10–11 for each version of TAB. There were no significant differences in the proportions of assessors from different staff groups scoring trainees on each version of TAB (Kruskal–Wallis test).

Over 400 assessors’ views were available for each version of TAB with the exception of TAB9, for which 89 views were available. The latter low questionnaire response rate was due to a clerical error: a batch of TAB9 forms were distributed without the appropriate questionnaire on the reverse side.

MSF scores using the different TAB forms

Table 2 shows the percentage of assessors scoring trainees as of ‘some concern’ or ‘major concern’ (TAB3 and 4), as ‘borderline or below expectations’ (TAB6) or as ‘unsatisfactory’ (TAB9). These descriptors all clearly suggest problematic behaviour. It can be seen that, for all forms of TAB, the domain ‘maintaining trust’...
had the least number of ‘concern’ or equivalent ratings, with 0.1%–0.7% of assessor responses.

Within the other domains, there appeared a clear trend with fewer ‘problem’ scores made by assessors using TAB9 than those utilising TAB3 and TAB4. Assessor responses using TAB6 were intermediate in their scores between those using TAB3 and TAB9: in the domain of verbal communication, 0.7% of assessors utilising TAB3 and TAB4 scored trainees as ‘below expectations’ as opposed to 0.1% of assessors using TAB9 and 2.4% of assessors using TAB3; for the teamworking domain, 1.4% of assessors using TAB6 scored trainees as ‘below expectations’ compared with 0.9% using TAB9 and 2.9% using TAB3.

Table 3 shows the percentage of assessors scoring trainees as ‘above expectations’ according to TAB form. TAB3 is excluded from this analysis as there is no ‘above expectations’ option in this version. A clear pattern appears, across all four domains, of higher numbers of assessors scoring trainees as ‘above expectations’ using TAB6 and TAB9 than using TAB4.

Analysis using the hierarchical loglinear model resulted in a model with a high level of fit ($\chi^2=9.1$, p=0.96) which included only two significant two-way interactions: (1) TAB rating scale version and rating (categorised as ‘problem’ vs ‘no problem’ scores) and (2) domain and rating (again categorised as ‘problem’ vs ‘no problem’ scores) ($\chi^2=28.1$, $\chi^2=30.7$, respectively, both p<0.001). This suggested that there was a highly significant association between these pairs of factors. In order to aid interpretation, the data were collapsed into two-way tables with these factors. Decreasing trends were noted in proportions of ‘problem’ scores for the TAB variants in the order: team working (1.8%), accessibility (1.7%), verbal communication (1.2%) and maintaining trust (0.4%), and for the TAB variants, TAB3 (2.0%), TAB4 (1.6%), TAB6 (1.2%) and TAB9 (0.5%). Put simply, this analysis suggests that the number of trainees identified with potential problems first, varies between the four different domains and second, is affected by the length of the rating scale on which the assessor scores the trainee.

Assessor views

Assessors were asked to give their views regarding the TAB process on a simple questionnaire on the reverse of the TAB form. The number of responses according to TAB form is shown in table 1.

Assessors were asked to agree or disagree (on a 6-point Likert scale) with the statement, “It’s difficult to tell the difference between some points on the ratings scale”. Responses show that higher proportions of those who used the two longer rating scales agreed with the statement. The percentages scoring 5 and 6 (ie, strongly agreeing with this statement) were 11% (TAB3) and 15% (TAB4), increasing to 23% (TAB6) and 21% (TAB9). Conversely, the percentages scoring 1 or 2 (ie, strongly disagreeing with this statement) were 50% (TAB3), 44% (TAB4), 37% (TAB6) and 38% (TAB9). However, if the whole scale was used, the difference in responses of assessors to this statement did not reach statistical significance (p=0.09, Kruskal–Wallis test).

We were also interested to know whether the different forms gave varying opportunity for assessors to give positive feedback to trainees doing well. The answer appeared to be that there were no significant differences (Kruskal–Wallis) in this respect. Percentages agreeing with the statement, “This form enabled me to congratulate good behaviour, where appropriate”, were 65%, 66%, 69% and 72%. Similarly, assessor responses to the question, “How effective is the form at providing worthwhile feedback?” were remarkably similar for all versions of TAB, with 48%–54% of assessors rating it ‘very effective’ and 3%–7% rating it ‘not effective’.

In response to the question, “How effective is the form in identifying trainees who may not be reaching the required professional standard?” the percentages scoring 5 and 6 (very effective) were 51%, 48%, 48% and 43% for assessors using TABs 3, 4, 6 and 9, respectively. These differences were again not statistically significant (Kruskal–Wallis).

DISCUSSION

Principal findings

In this study, we have attempted to explore the impact of rating scales on the scores given to a FY2 postgraduate medical trainee by assessors in a validated MSF assessment system. It has been found that: (1) longer rating scales are associated with a smaller proportion of trainees being identified as of potential concern; (2) longer rating scales are associated with a larger proportion of trainees being identified as above expectations; and (3) most assessors have no conscious problem using a variety of different rating scales.

Methodological issues

This is a difficult area to research robustly. One approach would be to get the same assessors to complete a variety of assessment forms on given trainees and then to investigate any differences in scores. However, it is likely that scores on one form will be influenced by scores given on another form. If assessors were asked to complete two or more forms with a suitable time gap, one could not differentiate differences arising from changes in assessor perceptions over time from differences arising from the form. Moreover, it may be difficult to get assessors to complete more than one assessment form, bearing in mind that the assessors are a mixture of co-workers of the trainee rather than, for example, teaching faculty. We therefore chose a study design

<table>
<thead>
<tr>
<th>Tab 2</th>
<th>Concerns scores by TAB variant, by domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TAB3, ‘some or major concern’</td>
</tr>
<tr>
<td>Maintaining trust</td>
<td>0.2% (1/537)</td>
</tr>
<tr>
<td>Verbal communication</td>
<td>2.4% (13/544)</td>
</tr>
<tr>
<td>Team working</td>
<td>2.9% (16/543)</td>
</tr>
<tr>
<td>Accessibility</td>
<td>2.4% (13/545)</td>
</tr>
</tbody>
</table>

Table 2: Above expectations’ by TAB variant, by domain

<table>
<thead>
<tr>
<th>Tab 3</th>
<th>‘Above expectations’ by TAB variant, by domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TAB4, ‘above expectations’</td>
</tr>
<tr>
<td>Maintaining trust</td>
<td>50.4% (370/734)</td>
</tr>
<tr>
<td>Verbal communication</td>
<td>42.4% (320/755)</td>
</tr>
<tr>
<td>Team working</td>
<td>51.0% (386/757)</td>
</tr>
<tr>
<td>Accessibility</td>
<td>44.1% (328/743)</td>
</tr>
</tbody>
</table>
in which assessors at different hospitals assessed trainees using versions of an MSF feedback form with different rating scales.

Our overall response rate in this study covered approximately 71% of trainees. Taking into account that, at the time of this study, MSF for FY2 doctors was not mandatory and that the denominator included trainees who could not take part because of absence for various reasons, we think this is an acceptable response rate and one from which it is reasonable to draw conclusions. It is possible the relatively high participation rate may be related to the brevity of the TAB. We acknowledge that participation is higher when it is mandatory.

**Interpretation of main findings**

With our approach, we found differences in assessments made using the different rating scales. There are various potential explanations for these differences. The first is that there were differences between trainees in the different hospitals, that is, those trainees in hospitals utilising TAB6 and TAB9 had higher levels of professionalism than those in hospitals utilising TAB3 and TAB4, as evidenced by the higher percentage of ‘concern’ ratings and lower percentage of ‘above expectations’ ratings in the latter settings. Clearly it is impossible to disprove this potential explanation, although there is no reason, in terms of selection method or institutional excellence, to think that this would be the case.

A second possible explanation lies with the assessors, either in their thresholds for acceptability and excellence or in their interpretation of the scales or descriptors. In terms of the former explanation, we have, in other work, found differences in assessors’ thresholds, but this has been in terms of professional group. Consultants have tended to score consistently more stringently than fellow trainees. However, in the current study, assessor groupings were similar throughout the participating hospitals. There is no reason that we can identify to suppose that assessors within the different hospitals would be more or less stringent than colleagues in neighbouring units. It is, however, possible that the assessors’ interpretation of ‘concerns’ is different to the interpretation of ‘below expectations’, which in turn is different from ‘unsatisfactory’. At the other end of the spectrum, ‘above expectations’ was a consistent category across TAB4, TAB6 and TAB9, yet differences in numbers scoring in this category were also seen. Moreover, that over 70% of assessments scored trainees as ‘above expectations’ using TAB6 and TAB9 does raise the issue of how assessors are using these descriptors. Whether it is the scale or the descriptor that pushes an assessor to a particular mark, it is clear that the 3-point scale and ‘concern’ descriptors of classical TAB seem to pick up more trainees with whom assessors are not completely happy than do the scales and descriptors as used in the Mini-PAT and Royal College of Physicians MSF tools. Further study would be needed to determine the cause and the respective parts played by descriptors and scale length.

The third potential explanation is that the rating scale itself affects the score given. With longer rating scales (those with more points on the scale), there seems to be a ‘right shift’ whereby trainees are scored more highly, with fewer ‘concern’ and more ‘above expectation’ ratings and this is proved by loglinear models. Further work, at least replicating this study, would be required to confirm or refute this possibility. If the rating scale does have the effect we have postulated, this has important implications for MSF in practice. If one imagines that the small percentage of trainees giving serious cause for concern, in terms of their professionalism, is contained within the lower end of scores allocated, then, in the case of TAB9, that would include trainees actually scored as ‘satisfactory’. Trainers would then have the challenge of tackling concern behaviour in trainees explicitly scored as ‘satisfactory’. Given that problem trainees at the severe end of the spectrum often lack insight, this could present a major challenge.

We were interested in the assessors’ responses to a simple questionnaire exploring their perceptions of the practicality and usefulness of the given model of TAB. Previous work has shown assessors to consider TAB to be a simple, practical and useful method of assessing professionalism. The current study appears to show that most assessors have no conscious problem using a variety of different rating scales, although more of those using the longer rating scales did seem to find it “difficult to tell the difference between some points on the ratings scale”.

It is interesting that no statistically significant differences were found in the responses to the question about the effectiveness of the form in identifying trainees who may not be reaching the required professional standard. This suggests that all TAB variants were equally able to perform this function.

**CONCLUSION**

This study provides evidence to suggest that the rating scale itself may impact on the assessment results in an MSF process. Although descriptions of points on scales may assist assessor judgement, this is an issue where it results in a problem trainee being labelled as ‘satisfactory’. If longer scales continue to be employed in MSF, consideration could be given to the use of a norm referenced lower proportion of trainees being considered for further support. A clear need for assessor training is suggested when MSF results indicate that over three-quarters of trainees are deemed ‘above expectations’.

**Main messages**

- The choice of rating scale may have an important impact on how assessors score trainees in the context of multi-source feedback to junior doctors.
- Longer rating scales seem associated with identifying fewer ‘concern’ trainees and more ‘above expectation’ trainees.
- Rating scale anchors alone may be inadequate to ensure robust judgements are made and recorded on multi-source feedback questionnaires.
- Assessor training remains a crucial component of effective workplace based assessment of junior doctors.

**Current research questions**

- What is the optimum rating scale to use in multi-source feedback (MSF)? (This also applies to other tools used in workplace based assessment.)
- How can assessors be facilitated in conducting robust workplace based assessments of trainees?
- How effective are current systems of MSF at identifying problem trainees?
- What is the outcome of trainees identified as ‘of concern’ by current MSF systems?
- How effective are current MSF systems at promoting professionalism among trainee doctors?
Key references


Competing interests None.

Contributors All authors contributed to the design and analysis of the study as well as the writing and review of the manuscript.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES


Online Archive

Visit our Online Archive — available back to 1925. Subscribers may access the entire archive freely. Non-subscribers have free access to all articles prior to 2006. A simple one-time registration is required that grants access to all the free archive content, across all of our specialist titles. To view or to register visit pmj.bmj.com.
Effect of rating scales on scores given to junior doctors in multi-source feedback

Andrew Hassell, Alison Bullock, Andrew Whitehouse, et al.

Postgrad Med J 2012 88: 10-14 originally published online November 3, 2011
doi: 10.1136/postgradmedj-2011-130010

Updated information and services can be found at:
http://pmj.bmj.com/content/88/1035/10.full.html

These include:

References
This article cites 15 articles, 2 of which can be accessed free at:
http://pmj.bmj.com/content/88/1035/10.full.html#ref-list-1

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/