Reviewer’s report

Title: Evaluation of physicians’ professional performance: An iterative development and validation study of multisource feedback instruments

Version: 1 Date: 13 January 2012

Reviewer: Julian Archer

Reviewer’s report:

Thank you for asking me to review this paper by Overeem et al. which I read with interest. The paper describes the psychometric validation of three MSF instruments modified from the PAR programme in Alberta for use with doctors in the Netherlands. The paper describes their initial development and early validation using a volunteer group of 146 doctors across 26 hospitals.

The paper is well written and clear overall. It is an interesting paper and provides some new insights but I have a series of concerns that I will discuss in this review.

The major compulsory revisions (to use the journal’s terminology) that need to be addressed are:

1. The review of the latest literature is insufficient and therefore the paper struggles to add new information about the use of MSF and patient feedback to the literature. While I accept that it is fundamental work for the instruments’ use in the Netherlands, the new or different areas are not brought out well. The paper starts correctly by placing the work in context but the literature quoted is somewhat dated. The authors describe MSF as ‘a relatively new tool’ (line 57), which is true of medicine, but not the wider literature. Early studies pre-date the Second World War. They quote Evans et al. and their assertion that there is a lack of evidence for MSF in the medical literature (line 69) but this is a 2004 paper which has been superseded by a growing body of publications since. Drawing on Lockyer’s review for the Royal College of General Practitioners would have been helpful as would have the latest paper from Campbell et al 2011 BMJ which used a more comprehensive modelling approach. There are important differences about the results compared to the literature but I wonder if this should not have been the focus of the paper rather than just in the discussion. I am thinking in particular of the correlation between patient and other feedback that is rarely seen.

2. I wonder about the IPR situation of the PAR programme and wonder about any issues surrounding the development of these instruments. I assume the authors have permission from the PAR programme to modify and use the instruments explicitly.

3. The analysis is conventional in its approach but does not follow expected reporting standards at all times. Crossley et al Pickin’ up good regressions Med Ed provides a useful model for good reporting practice including the need to
provide variance components etc.

4. I am concerned about a plagiarism issue with the passage: The CI, generated for the number of 'raters' that contributed to the individual 'physician' mean score, can then be placed around that score. This provides a measure of precision and, therefore, the reliability that can be attributed to each mean score based on the number of individual scores contributing to it (lines 197-200). This is referenced to Archer et al 2008 [23] in the text but is in fact verbatim from the earlier reference; Archer & McGraw 2010 [22]. This should be re-written or quote marked as verbatim.

The minor essential revisions (again to use the journal's terminology) are:

1. The removal of items if the 'unable to evaluate' rate more than 40% is very high. Lockyer et al 2006 Can J Med and others have used 15%.

2. The authors say that 'they could only use 80 percent of peer responses due to missing values on one or more items' (line 329) but the way in which missing data is dealt with is not described anywhere else. Is this different to 'unable to evaluate' scores?

3. There are assertions about stats signif relationships between the questionnaires but these are small (r=0.2-3).

4. Lines 319-321 assert that nesting 'produces more adequate results'. I am not sure that this is true unless the authors mean ‘better’ reliability. Wilkinson et al Med Ed 2008 would argue otherwise about the problems of nesting (in discussion p371)

5. Line 350-351 is unclear and requires some re-wording

Discretionary revisions might include:

1. Removing the line (345) for the use of such instruments in Asia as this may appear rather ‘marketing’ in nature.

2. The authors discuss the importance of feedback briefly (line 341-2) and concluded on the importance of this. Reference to the Miller & Archer BMJ 2010 review would be helpful here.

3. The use of the term ‘scale’ (line 42) in the abstract is misleading. Perhaps ‘factor’ would be better.

4. The authors might wish to consider analysis within the factors identified including reliability.

5. The authors do not comment on the potential problems of clustering within institutions or the use of different types of doctors. Differences have been demonstrated with different medical groups.

I hope that these comments are helpful and seen as constructive in the revision of this paper.

Julian Archer

**Level of interest:** An article of limited interest
Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I do not have any conflicts of interest