Reviewer’s report

Title: Inter-rater agreement in the scoring of abstracts submitted to a primary care research

Authors:

Alan A Montgomery (alan.a.montgomery@bristol.ac.uk)
Anna Graham (a.graham@bristol.ac.uk)
Philip H Evans (p.h.evans@ex.ac.uk)
Tom Fahey (tom.fahey@bristol.ac.uk)

Version: 1 Date: 15 Jan 2002

Reviewer: Ms Susan van Rooyen

Level of interest: not specified

Advice on publication: Other (see below)

General comments

This paper attempts an original piece of research. However, I am unhappy about deficiencies in the methods and over-generalisation in the conclusions.

Objective 1 - investigation of agreement between referees.

The study has examined agreement using only two referees, and in my opinion, this is its most serious weakness. The authors are trying to draw widely generalisable conclusions from a very restricted dataset.

Objective 2 - investigation of scores for accepted vs. rejected abstracts.

This section of the study is much more sound, and is also more easily generalisable. However, here again, I would see the use of only two referees restricts the generalisability of the results, although to a lesser extent.

Compulsory revision. Discussion of the deficiencies of the study, with suggestion/recommendation of what needs to happen in order that this might be addressed, either in this study or in the direction of future research.

Specific comments

Comments requiring compulsory revision

1. Title
I think there is a word missing from the title - perhaps it should read "Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference"
2. Abstract
a) Methods, line 6. I am assuming that the scores from each referee are averaged, and therefore this would more correctly read "Mean total scores", rather than "Total mean scores".

b) Results. Looking at the results given against the data in the Table, I do not understand the results given. According to the data in the Table, the ICC for originality and importance were both zero.

c) Conclusions. This statement could reasonably have been made simply from the original hypothesis, and not from evidence gained from this study. It should either be re-phrased to relate to the evidence gained from this study, or be added to, in order to emphasise the particular conclusions drawn from the evidence of this study. There are no conclusions deriving from the second objective of the study and something needs to be added concerning this.

Paragraph 2. It seems clear from the figure that the "mean score" refers to the average of the total score over the seven categories between the two referees, AM and AG. However, this needs to be more clearly explained in the text. I think it is also important to clarify how the intra-cluster correlation coefficients were calculated. I am assuming that the results are for correlations between reviewers on each of the seven categories scored.

I am also not entirely sure how appropriate it is to examine the mean score against the difference between reviewers, since these measures are not independent of one another.

4. Results.
The authors have mentioned "the three components that relate to study design" and to "subjective judgments about the importance and originality of the study". Categories 1 and 2 clearly fall under "subjective" and categories 3 and 4 fall under "objective". However, where were categories 5, 6 and 7 placed? This is not at all clear in the paper as it stands at present.

5. Discussion.
a) Paragraph 1. I think the conclusion is far too wide sweeping in the light of the fact that you only have the scores of two referees, and the abstracts have all been reviewed by the same two referees. To attempt to generalise is to draw far too grand a conclusion on the basis of the evidence so far. More discussion is needed on this point.

b) Paragraph 2. I do not understand the use of the word "inevitably" here, nor why the authors have this particular hypothesis that the scores will result from the reviewer's own research interests. This needs to be explained in more detail.

6. Conclusions.
The version of this paper which I downloaded via the Internet has no Conclusion section. It simply says: "Style tag for paragraphs within the Conclusions section". I therefore cannot comment. However, I refer to my comment concerning the Abstract conclusions.

Discretionary revisions

Discussion
1. "Abstracts accepted for the meeting had significantly higher quality scores, but with considerable overlap due to acceptance being determined by aggregation with other reviewers' scores". Have the data been analysed using only the scores from the two author-referees? Is there still an overlap?

2. I do not find it surprising that subjective components result in greater disagreement between reviews. However, I wonder if the authors have considered other criteria in determining "agreement"? My own experience is that different users of an instrument such as this will differ in part because some reviewers are "hawks" and some are "doves". For this reason, I feel that the conclusions drawn in this paper cannot be generalisable.

**Competing interests:**

None declared.