BMC Health Services Research

Research article

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

Alan A Montgomery*[1], Anna Graham[1], Philip H Evans[2] and Tom Fahey[1]

Address: [1]Division of Primary Health Care, University of Bristol, Cotham House, Cotham Hill, Bristol BS8 2PR, UK and [2]Institute of General Practice, School of Postgraduate Medicine and Health Sciences, University of Exeter, Barrack Road, Exeter EX2 5DW, UK

E-mail: 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

*Corresponding author

Published: 26 March 2002

BMC Health Services Research 2002, **2**:8

Received: 3 December 2001
Accepted: 26 March 2002

This article is available from: http://www.biomedcentral.com/1472-6963/2/8

Abstract

Background: Checklists for peer review aim to guide referees when assessing the quality of papers, but little evidence exists on the extent to which referees agree when evaluating the same paper. The aim of this study was to investigate agreement on dimensions of a checklist between two referees when evaluating abstracts submitted for a primary care conference.

Methods: Anonymised abstracts were scored using a structured assessment comprising seven categories. Between one (poor) and four (excellent) marks were awarded for each category, giving a maximum possible score of 28 marks. Every abstract was assessed independently by two referees and agreement measured using intraclass correlation coefficients. Mean total scores of abstracts accepted and rejected for the meeting were compared using an unpaired t test.

Results: Of 52 abstracts, agreement between reviewers was greater for three components relating to study design (adjusted intraclass correlation coefficients 0.40 to 0.45) compared to four components relating to more subjective elements such as the importance of the study and likelihood of provoking discussion (0.01 to 0.25). Mean score for accepted abstracts was significantly greater than those that were rejected (17.4 versus 14.6, 95% CI for difference 1.3 to 4.1, p = 0.0003).

Conclusions: The findings suggest that inclusion of subjective components in a review checklist may result in greater disagreement between reviewers. However in terms of overall quality scores, abstracts accepted for the meeting were rated significantly higher than those that were rejected.

Background

Interest in the peer review process and research aimed at determining the method of obtaining the best quality reviews has grown in recent years. [1] Checklists have been developed that aim to guide reviewers when assessing the quality of papers, but little evidence exists concerning the extent of agreement between two referees when evaluating the same paper. In addition, little is known about which dimensions of a checklist are likely to result in greater agreement between referees.

There were two aims of this study: (1) to examine inter-rater agreement of the quality of abstracts submitted to a primary care research conference (Annual Meeting of the

South West Association of University Departments of General Practice, Exeter 2000, UK), and (2) to compare the scores of abstracts accepted and rejected for the meeting.

Materials and Methods

Abstracts were anonymised and scored using a structured assessment comprising seven categories: (1) importance of the topic (2) originality (3) overall quality of the study design (4) appropriateness of the design used (5) achievement of aim (6) contribution to academic primary care (7) likelihood of provoking discussion. For comparison purposes, we have classified the assessment of categories 1, 2, 6 and 7 as more 'subjective' in nature, and categories 3, 4 and 5 as more 'objective'. Between one (poor) and four (excellent) marks were awarded for each category, giving a maximum possible score of 28 marks. Every abstract was assessed independently by two referees (AM and AG).

Agreement between referees was assessed using intraclass correlation coefficients (ICC), a chance corrected measure of agreement. [2] The ICC indicates perfect agreement only if the two assessments are numerically equal and is preferable to the more usual (Pearson) correlation coefficient. The crude ICC is lowered by any systematic differences between referees' scores. In terms of a plot of the two referees' scores, a line with a non-zero intercept will further lower the ICC irrespective of any disagreement, represented by deviation of the slope of the line away from unity and scatter around the line. In a further analysis, this effect was investigated by subtracting the mean difference for each component from the higher of the two referees' scores. The ICCs were then recalculated, giving estimates of agreement corrected for both systematic differences and chance. There are no universally applicable standard values for the ICC that represent adequate agreement, but the following convention is used here to aid interpretation: ICC <0.20 'slight agreement'; 0.21–0.40 'fair agreement'; 0.41–0.60 'moderate agreement'; 0.61–0.80 'substantial agreement'; >0.80 'almost perfect agreement'.

Scores from referees from three different institutions were summed to give each abstract an overall score. Abstracts were ranked by this overall score and the top 45 were accepted for oral presentation at the meeting. Of the 52 abstracts refereed by AM and AG, mean total scores of those accepted and rejected for the meeting were compared using an unpaired t test.

Results

Chance corrected agreement between the two referees' scores measured using crude ICCs was greater for the three components relating to design and execution of the study (Table 1: items 3 to 5) compared to those relating to more subjective elements of the abstract (Table 1: items 1, 2, 6, 7). After adjustment for systematic differences in referees' scores, ICCs for items 3 to 5 remained highest, demonstrating fair to moderate agreement.

A total of 76 abstracts were submitted for the meeting. Of 52 received by the authors for assessment, 26 were accepted for oral presentation (Table 2). Abstracts accepted for the meeting had a significantly higher mean score than those that were rejected (95% CI for difference 1.3 to 4.1, p = 0.0003) (Table 2).

Discussion

This study has shown that when using a structured assessment form, two independent reviewers were more likely to agree on design or methodological components of a checklist than on subjective components of abstracts submitted for an annual research meeting. Abstracts accepted for the meeting had significantly higher total scores, but overlapped considerably with rejected abstracts. This was due to acceptance for the meeting being determined by an overall aggregate of scores awarded by referees from three institutions.

Table 1: Inter rater agreement between two referees for 52 abstracts submitted for a primary care research conference

Component	Mean difference (AG minus AM)	Crude ICC	Adjusted ICC
1. importance of the topic	0.71	0	0.24
2. originality	0.27	0	0.01
3. overall quality of the study design	0.40	0.30	0.40
4. appropriateness of the design used	0.17	0.40	0.41
5. achievement of aim	-0.12	0.44	0.45
6. contribution to academic primary care	0.25	0.20	0.25
7. likelihood of provoking discussion	0.15	0.22	0.24
Overall score	1.85	0.31	0.41

ICC = intraclass cluster coefficient

While the subject of inter-reviewer agreement on different components of a checklist is relatively under-researched, some previous studies offer support for our finding that agreement is better when reviewers can be more objective in their assessments. Among a group of reviewers asked to rate a series of review articles, agreement on scientific quality of the papers was very high (60% of ICCs > 0.7) both within and between groups with varying levels of research training and expertise. [3] All 10 dimensions of the checklist that reviewers rated could be regarded as objective. Divergent reviewers have been identified in a study comparing an overall rating score that indicated a recommendation to publish rather than individual dimensions of a review checklist. [4]

This study does have limitations. Importantly, we assessed agreement between only two reviewers on a relatively small number of abstracts. This could be addressed by having more abstracts assessed by a greater number of reviewers. However the study was conducted pragmatically within the time and administrative constraints of a small annual scientific meeting rather than submissions to a journal over an extended period. Another limitation is that the reviewer checklist was constructed prior to conceiving the study. If future meetings are to be used to investigate the content of structured reviewer assessments, such checklists should be constructed with specific hypotheses in mind.

Table 2: Summary statistics of abstracts accepted and rejected for oral presentation at a primary care research conference

	N	Mean score	SD	Range (possible 7 to 28)
Accepted	26	17.4	2.7	11 to 21.5
Rejected	26	14.6	2.3	11 to 19.5

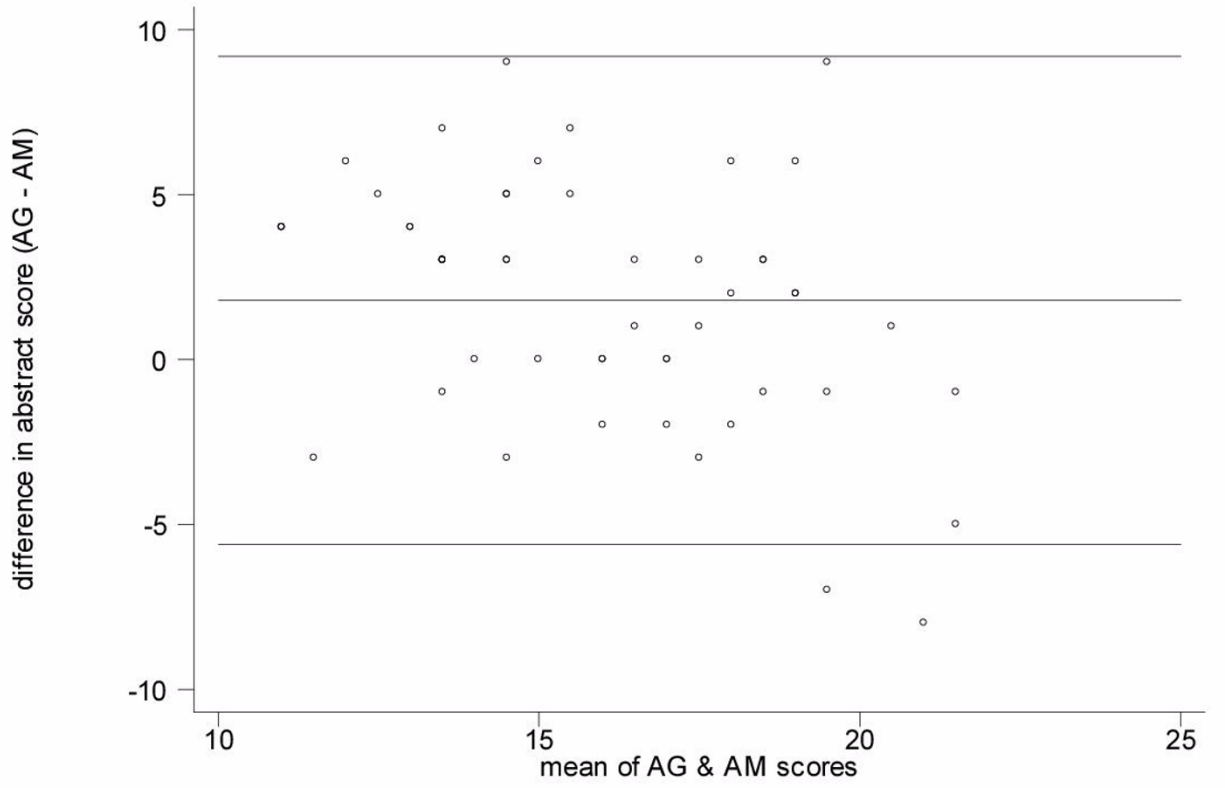

Figure 1
Difference between referees' scores versus mean score

Characteristics associated with good peer review are age under 40 years and training in epidemiology or statistics, [5] characteristics that applied to both reviewers in the present study. Structured assessment forms that ask the reviewer for their opinion of a paper's interest, originality or likelihood of provoking discussion may be more likely to result in scores that reflect the reviewer's own research interests. This is not necessarily a criticism – it is perhaps only natural that individuals will differ in their opinions of how interesting they find, and think others will find, a particular paper. It is interesting that the two components with the lowest agreement, importance of the topic and originality of the study, both require more knowledge about a specific subject area than either of the other two subjective questions. Journal editors and meeting organisers should be aware that including subjective components in review checklists may result in greater disagreement between reviews.

Conclusions

This study provides some evidence that inclusion of subjective components in a review checklist may result in greater disagreement between reviewers. An interesting area for further research would be to investigate the effects of attaching different weights to subjective and objective components of a checklist, or to exclude subjective components altogether from overall quality scores and simply use them a guide to acceptance or rejection.

Competing interests

None declared.

Acknowledgements

We would like to thank Mrs Sylvia Smith for collation of all referees' responses. During the period in which this work was conducted, AM was supported by a UK Medical Research Council Training Fellowship in Health Services Research. TF is supported by a UK NHS R&D Primary Care Career Scientist Award.

References

1. Smith R: **Peer review: reform or revolution.** *BMJ* 1997, **315:**759-760
2. Streiner DL, Norman GR: *Health measurement scales: a practical guide to their development and use.* 1989
3. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, *et al*: **Agreement among reviewers of review articles.** *J Clin Epidemiol* 1991, **44:**91-98
4. Siegelman SS: **Assassins and zealots: variations in peer review.** *Radiology* 1991, **178:**637-642
5. Black N, van Rooyen S, Godlee F, Smith R, Evans S: **What makes a good reviewer and a good review for a general medical journal.** *JAMA* 1998, **280:**231-233

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6963/2/8/prepub