Author's response to reviews

Title: Reviewer Agreement Trends from Four Years of Electronic Abstract Submissions

Authors:

Brian H Rowe (brian.rowe@ualberta.ca)
Carol Spooner (cspooner@ualberta.ca)
Sandra Blitz (SandraBlitz@cha.ab.ca)
Trevor L Strome (tstrome@wrha.mb.ca)
Eric Grafstein (egrafstein@providencehealth.bc.ca)
Andrew Worster (aworster@rogers.com)

Version: 3 Date: 23 January 2006

Author's response to reviews: see over
COMENTS LETTER

Response to Reviewer 1 – Karen Shashok

Section 1 – General
Research on peer review methods is necessary, and strategies to make the process more efficient and more reliable are needed. This retrospective analysis of data obtained over 4 years with a sample of abstracts for a scientific society's annual conferences shows interesting evidence that reviewers are more likely to agree on objective, clearly definable criteria (methods, statistics) and on the overall rating of a given abstract than on more subjective criteria (the judgment of which is probably influenced by many uncontrollable reviewer-dependent factors such as education, training, experience, school of thought, motivation to review well, and reviewing skills).

Author response: Thank you for this summary and comment.

2. The Introduction and Discussion sections contain material that is not relevant to the objectives of the study or that is speculative and not based on the evidence obtained in this study. This problem can be dealt with by deleting some material as suggested below so that the revised paper focuses on the objectives, data, and conclusions that can legitimately be drawn from the findings. This will make the paper easier for readers to understand.

Author response: We have endeavored to revise the paper in the manner suggested by the reviewer. The sections are condensed and we agree this makes it easier for consumers to read.

3. The data are presented confusingly in the text. Specific places were confusion arises are noted below.

Author response: We revised the presentation of data throughout.

4. It would be reassuring if the authors could provide references to reliability and validation studies for the instrument they used to evaluate the abstracts.

Author response: Unfortunately, there are no reliability and validity papers that we could find on the topic. These forms have been used by the organization and iteratively changed over time to reflect comments.

Section 2 – Major Compulsory Revisions
1. Background. The first 2 paragraphs of this section are not directly related to the study and can be deleted.

Author response: We believe that information in paragraphs 1 and 2 of the background section serve to present the context of the paper. We have condensed the most salient points to one paragraph.

2. Background. The third paragraph ("The traditional methods of accepting abstracts...") and fourth paragraph ("Many of the advantages...") tend to confuse administration of the reviewing process (sending and receiving files, record-keeping) with the evaluation of the scientific content. These are two different processes that need to be considered separately. There seems to be little evidence as yet that using e-mail instead of snail mail ("electronic peer review") improves the quality of scientific critiques, although it does shorten the time to decision by saving the time manuscripts and reports would otherwise spend in snail mail.

Since it is probably not necessary to explain the advantages of e-mail for administering peer review systems, these two paragraphs can be deleted.
**Author response:** Paragraphs 3 and 4 of the background section have been removed to keep focus of the paper on reviewer agreement.

3. Background. A better place to start this section would be with paragraph five, "The ability to effectively identify..."

**Author response:** This paragraph has been moved to the second in the Introduction section.

4. Background. In the paragraph that begins "Despite the variabilities..." it would be useful to clarify that 1) the instrument investigated in this study was designed for conference abstract peer review rather than journal article peer review, and that 2) the aims of these two processes are different, since the purpose of conference abstract review is to screen studies and decide which will be included on the conference program and which will be excluded, whereas the purpose of journal article abstracts is to reflect the content of the full text accurately. So the criteria used to judge abstract quality are somewhat different in each case.

**Author response:** As suggested, we have added a brief description of the differences between abstract reviews for conferences and manuscript submissions for journals. The fact that our instrument is based designed for abstract reviewing is added in the methods section.

5. Background. "Conversely, higher inter-rater agreement indicates a more reliable system that is less likely to be influenced by bias." This statement should be qualified; higher reliability does not guarantee that the system is less susceptible to bias. The biases may operate in highly reliable or reproducible ways, i.e., the system might be reliably biased” or "reliably unfair”.

**Author response:** As suggested, we have added a section on systematic/reproducible bias to the last paragraph of the background section.

6. Methods. Has VS Review been tested in other samples of abstracts in other populations of reviewers in other disciplines? Is there evidence available of the reproducibility of the evaluations?

**Author response:** The evaluation criteria used by VS Review are customizable for the client’s requirements. CAEP is the only group using this particular evaluation criteria set, so comparison to other groups’ inter-rater reliability, while interesting, unfortunately would not add anything meaningful to this study. This is why a longitudinal evaluation of the CAEP reviewers is the most valid comparison currently available.

7. Methods. Readers will need more information on how the criteria used in VS Review are described to reviewers and how reviewers are asked to evaluate each criterion. For example, are the criteria presented as a checklist with only 2 response options (yes/no, present/absent), or are the criteria rated on a Likert-like scale of 3, 4 or 5 points, or on a continuous scale from 1 to 10?

**Author response:** More details (and a table) have been added regarding the criteria development and scoring methodology.

8. Methods. "...there are three sources of variance in scores that impact reviewer agreement: 1) abstract effect; 2) reviewer effect and; 3) abstract-reviewer effect.” This is only one of many possible approaches to analyzing sources of variance, so the sentence should be edited to reflect this.

**Author response:** The sentence has been modified to reflect that our approach is only one of many possible approaches to analyzing sources of variance.

9. Methods. "The reviewer-abstract effect is a non-systematic positive or negative reaction a reviewer might have to a given abstract that results in the reviewer assigning a higher or lower
score than would normally be assigned." It would be helpful here to explain how this effect is different from "bias" (in this case, individual, reviewer-related bias).

**Author response:** A section has been added to discuss how the effect is different from bias. The authors suspect that topic preference may lead to both up-regulation and down regulation by the same reviewer, so the effects should cancel out.

10. Methods. Analysis. The changes in criteria in 2003 should be explained briefly in this section, since when readers come to Table 1 they will wonder why, starting in 2003, 1) there are two Methods criteria instead of one, 2) Impact was scored as a maximum of 4 rather than 3, 3) Originality was not evaluated in 2001-2002, and 4) Conclusions and Recommendations were not evaluated in 2003-2004.

**Author response:** The changes to the criteria were based on the comments and feedback from reviewers to help clarify and remove ambiguity from the criteria. We have added this information to the text.

11.1. "The mean final score...P=0.229)." This information is not relevant to the aims of the study, and can be deleted.

**Author response:** As suggested, we have deleted this sentence.

11.2. "The 95% CI of the ICC values crossed the 0.40 ("moderate agreement") line in all cases..." Please rephrase this as something like "The 95% CI of the ICC values were 0.41 or higher ("moderate agreement") in all cases..."

**Author response:** As suggested, we have rephrased this sentence.

11.3. Paragraph that begins "Table 3 illustrates..." This paragraph is confusing. Please rewrite it to make clear when you are comparing the ICC for certain criteria across the 4-year study period, and when you are comparing the ICC for all criteria within a given year.

**Author response:** As suggested, we have re-worded the paragraph to make it clearer.

11.4. "...and the most variation occurred for Methods I (variation = 0.27)." According to Table 3, this value (0.57 minus 0.26) should be 0.31. Please amend the data in the text or in Table 3 as necessary.

**Author response:** Correction have been made in the text.

11.5. "...the 95% CI of several criteria ICC values cross the 0.40..." What year does this refer to?

**Author response:** Clarifications have been made to the text.

11.6. "The 2003 criteria ICCs ranged from 0.25 (Hypothesis) to..." Please edit this to read "(Hypothesis and Originality)" in accordance with the data in Table 3.

**Author response:** The edit has been made.

11.7. "The ICC was greater than 0.40 after 2003." Does this refer to the overall ICC for all criteria, and do you mean that it was greater than 0.40 only in 2004?

**Author response:** This comment was erroneously placed in that location and has been removed.

11.8. "A system was devised..." The justification for generating two rankings (one for 2001-02 and one for 2003-04) is clear but the system devised for ranking within blocks seems more complicated than necessary. Recalculating the ICC for each criterion within each block simply as the mean of the two ICCs for each year yields slightly different results for the Block 1 ranking in Table 4 for positions 3, 4 and 5. The top criteria in terms of reviewer agreement based on the rankings for 2001-2002 with this simpler approach are Statistics (1), Overall (2), Recommendation (3) and Methods (4). Is the method you used more statistically robust than just calculating the mean ICC for the 2 years to rank them?
**Author response:** The authors agree that the ranking system was unnecessarily complicated for the descriptive purposes of determining the criteria that resulted in the best reviewer agreement. Table 4 has been deleted and the text has been modified to reflect that simple averages for the rankings for 2001-2002 and 2003-2004 have been calculated.

12. Results, last paragraph. Readers would be curious to know if you can offer any explanation for the difference in ranking between Block 1 and Block 2, since the changes in the criteria used in the evaluation might not be able to account for all of them.

**Author response:** At this point, we have not formulated hypotheses regarding why there was a change in ranking between the two blocks. Every year there is some "jockeying" for position between the criteria, so a difference in ranking between blocks 1 and 2 is not surprising.

13. Discussion, second paragraph. "The ICC for Methods I increased substantially for 2003 and 2004 as compared to 2001 and 2002, despite the exact same criteria being used over all four years." This statement will surprise readers since in 2003-04 the Methods criterion appears to have been split into two elements: Methods I and Methods II. Confusion can be avoided by explaining in the methods section that a second Methods criterion was added in 2003 (see point 10 above).

**Author response:** Sentence has been reworded, and in conjunction with additional clarification regarding the changes in criteria (see above), should be clearer.

14. ...The authors should keep this in mind and not overemphasize the importance of surrogate measures of the quality of the science reported in abstract. The VS Review instrument is meant to make abstract selection cheaper and faster, but whether it also makes the outcome of the competition more reliable (only the best reports are selected and only poor reports are rejected) is still open to question. The reviewer agreement levels reported in this study are very modest, even for the more objective criteria. This suggests that the more subjective criteria were not understood by the reviewers, or were understood to mean different things. A better definition of the more subjective criteria, as the authors note, might improve agreement levels.

**Author response:** The paper has been modified to better reflect the authors’ desire to find sources of bias in the overall system, not to find statistical manipulations to make the reviewer agreement look better.

**Section 3 – Minor Essential Revisions**

1. Running title. This could be edited to something like "Agreement trends in conference abstract peer review".

**Authors response:** The running title has been modified to more clearly identify we are reporting conference abstract peer reviews.

2. Abstract, Methods. "design-specific form". Please reword this to explain what "design-specific" means.

**Authors response:** this meant that the review form was displaying the appropriate review criteria for a given study type design (i.e., RCT). This has now been clarified in the text.

3. Abstract, Conclusions. "Methodological criteria descriptions resulted in..." Please reword this; it is too condensed to communicate the meaning clearly.

**Authors response:** This has been re-worded in the text.

4. Abstract, Conclusions. "In future abstract competitions...inter-rater agreement." This sentence can be deleted since it is speculative, and in a sense misleading. Alternatively, the sentence could be edited to something like "...defining criteria more objectively so that reviewers can base their responses on empiric evidence..."
**Author response:** The conclusion section of the abstract has been re-worded to improve the clarity of the conclusions resulting from this study.

5. Background, last paragraph, point 1. The specific instrument analyzed in this report, and the fact that the study is about conference abstracts (as opposed to journal article abstracts) should be mentioned briefly here so that readers understand that the study is about a specific instrument and not about abstract peer review in general.

**Author response:** We have clarified in the text that this study is looking at conference abstract evaluation based on a tool developed by CAEP members and implemented in an electronic format via the VS Review software.

6. Background, last paragraph, point 3. This point should be deleted since suggestions for improvements and future studies are best saved for the Discussion section.

**Author response:** Point 3 has been removed.

7. Methods, Review System Overview. "...while conflict-of-interest declarations on the reviewers' part are encouraged if discovered while reviewing the abstract." Please rewrite this sentence to remove the ambiguity caused by "while" and "if discovered". Something like "...and conflict-of-interest declarations on the reviewers' part are requested if such conflict is discovered during abstract review" might be clearer.

**Author response:** This sentence has been re-worded.

8. Methods, Review System Overview. "When measuring the reliability of review criteria..." Please rewrite this sentence to remove the ambiguity caused by "When measuring". Something like "In measurements of the reliability of review criteria..." might be clearer.

**Author response:** This sentence has been re-written.

9. Methods, Review System Overview, last paragraph. "This is done primarily through carefully crafted scoring criteria, blinded reviewing, and removal of conflicts of interest." Conflict of interest cannot be removed once it arises, but it can be declared so that readers are aware of it. Perhaps "detection" or "declaration" would be a better word than "removal" here.

**Author response:** This section of text has been we-worked to improve clarity.

10. Methods, Analysis. "...only if assessments are numerically equal." Please rewrite this phrase for the benefit of readers who are not familiar with the ICC or the methods in references 10 and 11. Does this phrase mean "only when the number of observations or elements to be analyzed is the same"?

**Author response:** This section of text has been we-worked to improve clarity.

11. Methods, Analysis. "...and Kappa (another measure of rater agreement based on two raters)".

This phrase can be deleted since the kappa index is not relevant to this study.

**Author response:** although the Kappa index is not used in this study, the authors believe that more readers will be familiar with Kappa than with ICC, therefore mentioning Kappa in this analogous fashion will provide a tool to which they can relate.

12. Discussion, first paragraph. "First, agreement between reviewers on total score can be considered only "moderate" since the total score ICC exceeds the 0.40 value in every year except 2001." This sentence is a convoluted way to express this information. It would be easier on readers to write something like "agreement between reviewers on total score should be considered only "moderate" since the total score ICC was between 0.21 and 0.40 (fair) in 2001 and between 0.41 and 0.60 (moderate) in 2002, 2003 and 2004.

**Author response:** This section of text has been we-worked to improve clarity.

13. Conclusions. This section contributes little that is novel or based on the evidence reported in the manuscript, so it can be deleted.
Author response: We respectfully believe that the conclusions section provides a way to “wrap up” the paper, but have greatly condensed the section to include only the most relevant points.

14. Table 1 heading. This should be edited to something like "Abstract review criteria used, and maximum point values, in 2001-2002 and 2003-2004 conference abstract competitions.
Author response: The heading has been changed to reflect this recommendation.

15. Table 3 heading. This should be edited to something like "Variation in reviewer agreement for different criteria used to review conference abstracts. Criteria-level intra-class..."
Author response: The heading has been changed to reflect this recommendation.

16. Table 4 heading. This should be edited to something like "Review criteria ranking by reviewer agreement in each 2-year block"
Author response: This table has been removed from the paper altogether to enhance clarity; its findings have been summarized in the text.

17. References. This section should be checked carefully to correct misspellings (ref. 16 and possibly elsewhere), typos (for example, ref. 13, last author's initial missing) and errors in journal abbreviations (refs. 12 and 13). The ampersand before the last author's name should be omitted throughout the reference list.
Author response: The references have been re-formatted.

Section 4 – Discretionary Revisions

1. Methods, Review System Overview. "For example, review criteria for statistics and methodology differ between randomized controlled trials and qualitative studies.” Since the most notable levels of reviewer agreement in this study were found for methods-related criteria, it would be interesting to compare the results for abstracts of RCTs and abstracts of qualitative studies analyzed separately.
Author response: The authors agree that this comparison would be interesting. However, since there are rarely any submissions of qualitative studies to CAEP annual meetings, any conclusions drawn from such comparisons would be of limited value due to weak statistical power.

2. Results. It is interesting that for Block 1, the rank (Table 4) for Recommendation (3/9) was so far removed from that for Conclusion (7/9). Was the decision to omit both these criteria in the 2003-04 evaluations a consequence of this disparity in criteria that might be expected to be directly correlated?
Author response: This entire analysis was done retrospectively after the 2004 conference. The changes in the criteria, however, were the result of suggestions to enhance clarity and fairness received from reviewers after the 2001 and 2002 conferences before this quantitative data was compiled.

Response to Reviewer 2 – Liz Wager

Section 1 – Major Compulsory Revisions

1. It would be nice to have a more focused account about what is already known on the subject of reviewer agreement in the abstract. I felt this (and the Methods, 2nd para, p8) contained too much information about electronic submission, which was not really the focus of this study. Although electronic submission made the data collection feasible, it didn't seem fundamental to the study, so I felt the discussion about its merits was not particularly helpful (1st full para on p6). I would have preferred to see more detail about the review criteria, which are hard to judge from the short titles (eg Methods I and Methods II).
Author response: We have revised the introduction/background text in an attempt to make it more focused, we have reduced the emphasis on the computerized nature of the system and added more details about the review criteria themselves.
2. Given that BMC is an e-journal, I'd encourage the authors to append a review form (e.g. with criteria for RCTs -- I realize that you don't want to present all the different variants) as an appendix. This would also help readers judge for themselves which criteria are more or less objective.

**Author response:** We have added a figure that shows the complete review criteria for one of the submission types.

3. I would like to see more discussion about what happened in 2003 and why the authors think this seemed to change both the mean scores and the degree of agreement. Do they think this was mainly due to the change in criteria? If so, why was the agreement not maintained in 2004.

The authors should discuss the possibility that the effect was either a random variation, or a non-specific effect of change. Any change in the criteria might make reviewers read them more carefully and therefore be more consistent, and this could explain why the effect 'wore off' the following year. For this theory to be true we'd also need to understand how many reviewers had reviewed in previous years and how many were 'new'. There is good evidence that any change can have effects (eg changing light levels alters factory production, but it doesn't matter whether the change is an increase or a decrease, it's simply the change that matters) -- this needs to be discussed.

**Author response:** We believe it is possible that the apparent peak in reviewer agreement is due to the growing experience and competence of the pool of regular CAEP reviewers (75-80% overlap each year). The authors believe that the apparent drop in reviewer agreement after 2003 is due to the addition of several inexperienced reviewers to the pool for 2004. This inexperience might have lead to less consistency in the rankings of their assigned abstracts.

### Section 2 – Minor Compulsory Revisions

1. The abstract gives the impression that all abstracts were rated by just 3 reviewers (1st line of methods) -- I suggest this should be rephrased to read 'and each assessed by three (out of **) volunteer CAEP reviewers'.

**Author response:** The text has been re-worded to clarify that there were not just three CAEP reviewers.

2. The total number of reviewers should also be mentioned in the full text (somewhere in the Results section).

**Author response:** We have added the total number of reviewers for each year in the study.

3. Overall, the MS is well presented, but I spotted a few typo's: Ref 9: should be Intern Med (not Inern) Ref 16: should be judgement (not judement)

**Author response:** The typographical errors have been corrected.

4. A phrase on p.13 confused me. You state 'the total score EXCEEDS the 0.40 value' whereas I think you mean it crosses (but goes below) this.

**Author response:** The results sections have been re-worked to enhance the clarity of the information / data presented.

### Section 3 – Discretionary Revisions

1. One advantage of a database over several years is that presumably you can build up data about individual reviewer characteristics. Although this is mentioned in the discussion as a topic for further research, I would very much have liked to have seen something about the effect of individual reviewer characteristics on review scores. The definition of the types of variation (abstract / reviewer / abstract-reviewer, bottom of p9) is nice and clear (although maybe it would fit better in the Introduction than the Methods), but leads the reader to expect something about
reviewer variation. Is it possible to know, for instance, whether some reviewers scored consistently harsher or more leniently than others (i.e., were there 'assassins' and 'zealots')?

**Author response:** This is a very interesting topic and will warrant further investigation in a future paper; we believe this is beyond the scope of the original objectives of the study.

2. I would have liked more information / discussion about the 6-point item for 'overall' impressions. At first sight this seems very subjective, yet it got quite good agreement. If that's the case, one might argue that getting reviewers simply to use a 6-point subjective score would be just as good as the 'objective' assessments. It would therefore be interesting to know the extent to which the score for 'overall' correlated with the total score for each abstract. This would also help to validate the scoring system.

**Author response:** This is a good point. Based on this suggestion, we found a Pearson correlation $r$ of 0.91 ($p<0.05$), indicating a very strong relationship. This has correlation value has been added to the text.