Reviewer’s report

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

Version: 1 Date: 1 January 2006

Reviewer: Karen Shashok

Reviewer’s report:

General

1. 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 judgement 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).

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 focusses on the objectives, data, and conclusions that can legitimately be drawn from the findings. This will make the paper easier for readers to understand.

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

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.

-----------------------------------------------------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

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

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.

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

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.

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".

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?

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?

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.

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).

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.

The addition of a second Methods criterion should be justified briefly; otherwise some readers may be confused when they read the first sentence in the second paragraph of the Discussion, as noted in point 13 below.

11. Results. The text should be revised in a number of places to make the content easy for readers to assimilate:

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

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..."

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.

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.

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

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.

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?

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?

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.

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).

14. Discussion, second paragraph. "The Committee felt that this approach would reduce variability in the scoring." Without more information about the structure of the criteria and validation of the abstract evaluation instrument used, this statement may raise concerns about the ultimate aim of the VS Review instrument: to select "good" abstracts and reject "poor" abstracts with an acceptable level of sensitivity and specificity, or to yield scores with as little variability as possible for as many criteria as possible (putting the cart before the horse, so to speak). This concern is reinforced by the final sentence in the Discussion, "...and statistical methods to mitigate such effects...", which suggests that the authors (or possibly the peer reviewers of the abstracts and the CAEP) may be focussing on improving reviewer agreement statistics so much that there may be a temptation to lose sight of the ultimate purpose of the abstract competition: to identify science that is good enough to be reported at the annual meeting.

The aim of this study (and peer review research more broadly) should not be to find statistical manipulations to mitigate variability, but to design better evaluation instruments so that variability will be lessened. Refining the more subjective criteria "to incorporate less ambiguous guidelines or parameters" is a better approach than "modifications to the scoring sheets to reduce weightings of subjective criteria and increase weighting on more objective criteria". This latter would indeed improve inter-rater agreement statistics but would probably not improve the reliability of the actual outcome of the peer review process.

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.

---

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

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

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

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

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..."

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.

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.

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.

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.

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.

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"?
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.

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.

13. Conclusions. This section contributes little that is novel or based on the evidence reported in the manuscript, so it can be deleted.

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.

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 intraclass..."

16. Table 4 heading. This should be edited to something like "Review criteria ranking by reviewer agreement in each 2-year block"

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.

Discretionary Revisions (which the author can choose to ignore)

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.

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?

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

I declare that I have no competing interests.