IZA07149

Achieving Quality Reviews

John Izard
School of Education, RMIT University, Melbourne, Australia
<john.izard@rmit.edu.au>

At the AARE conference in Adelaide in 2006 a paper exploring the peer review of papers for the annual AARE conferences from the perspective of the reviewer was presented. The paper described common elements over a period of more than a decade including problems of research design, failure to validate assessment strategies prior to their use in research, inappropriate measures of learning, claiming that learning has occurred in spite of the evidence collected, inadequate consideration of competing plausible explanations for treatment effects, inappropriate application of statistical significance, and failure to indicate the magnitude of effects.

This paper takes a similar perspective in elaborating on assumptions made by journal editors about papers submitted for publication, papers accepted and rejected, advice given in justification of acceptance or rejection, and about quality control of the review process. The paper uses examples of interpretation of statistical significance, and the magnitude of effects to illustrate the issues. It questions whether current review procedures have adequate safeguards to avoid bias, and suggests improvements that may make such review procedures more transparent.

Keywords: Assessment and Measurement

Context

At the AARE conference in Adelaide in 2006 a paper entitled ***Peer review: Hits and misses*** (IZA06033) was presented. It explored the peer review of papers for the annual AARE conferences from the perspective of the reviewer. Since journal editors use a variety of similar procedures for papers submitted for publication it was judged appropriate to compare problems in peer review for AARE conferences with similar problems in peer review for journal publication, including the *Australian Educational Researcher*.

This paper commences with a tabulation of perceived problems in conference papers submitted for peer review over a period of more than a decade (see Table 1), and a consideration of the problems faced by reviewers in judging the quality of proposed papers for AARE conference (see Table 2). Similarities and differences between AARE conference refereeing and journal refereeing are discussed and implications for quality refereeing are presented.

Assumptions made by journal editors about numbers of papers submitted for publication, about the reporting of numbers of papers accepted or rejected, and about quality control of the review process are presented and discussed. The range of information required of referees to justify their decision is outlined, and the relevance of this information to improving future proposals is addressed. The usefulness of any

advice given in justification of acceptance or rejection is illustrated through examples of interpretation of statistical significance, and the magnitude of effects.

Comparison of AARE Conference and Research Journal Referee Procedures

It should be noted that review of abstracts to decide who will present at a conference is not usually regarded as peer review, since this is a mechanism to exclude proposals that are not relevant to the purpose of the conference, and proposals that are commercial in the sense of providing advertising for a product or process. In the context of this paper, consistent with AARE usage, the term *peer review* refers to a procedure where academics with expertise in the relevant study area *referee* a proposed full paper by one or more authors to judge whether it is of sufficient quality to be presented as a refereed paper and acknowledged as such. The system operated by the AARE Office from 1999 to 2007 is described by Peter Jeffery in his paper *AARE's Conference full papers academic peer refereeing processes* (JEF07611).

Perceived problems in **conference papers** submitted for peer review over a period of more than a decade are shown in Table 1. These problems include a multitude of issues relating to flaws in methodology, flaws in interpretation, inappropriate applications of judging statistical significance, and a failure to consider the magnitude of effects.

For example, if a researcher claims to have a **population**, a statistical significance test is **not** appropriate. [Statistical significance tests are used to infer a population statistic from a (simple or complex) random (representative) sample from the population.] The differences obtained are the actual differences that should be reported. See Thompson, B. (1999) 'Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.' Invited address presented at the annual meeting of the American Educational Research Association, Montreal. [http://acs.tamu.edu/~bbt6147/aeraad99.htm or www.coe.tamu.edu/~bthompson/aeraad99.htm]

Where a statistical significance test is appropriate, certain assumptions must be respected or the conclusions will not be valid. For example, claiming that one has a simple random sample implies that every potential participant had an equal opportunity of being selected in the study. Published statistical tables (and commonly used software like SPSS) assume simple random sampling. But using simple random sample assumptions with complex samples such as intact classes within selected schools could lead to more "significant" results than are warranted. In his University of Melbourne M.Ed. thesis research (1975), Ken Ross used different types of samples from a population. His research showed that gross errors of interpretation resulted if decisions based on simple random sample assumptions were used with complex random samples. For information on these issues, see module 3 [Ross, K. (2005). Sample Design for Educational Survey Research: Module 3. Paris: International Institute for Educational Planning (UNESCO).] in the International Institute for Educational Planning (IIEP) Training Modules in "Quantitative Research Methods in Educational Planning" available from http://www.sacmeq.org/training.htm or http://www.unesco.org/iiep .

Note that failure to reject the null hypothesis does *not* mean that there *is no difference*. Logically it means we have *no evidence* of a difference. We may have a true difference but the *power* of our statistical test to detect that difference may be inadequate. Or there may not be a difference. *We just do not know*.

There are other misconceptions about testing for statistical significance. Researchers who refer to a statistically significant result as "a reliable difference", (meaning one that is replicable) are in error. The statistical significance level does *not* indicate the replicability of research data. Further, the statistical significance level does *not* provide an index of the importance or size of a difference or relationship. A difference significant at the .001 level is *not* theoretically (or practically) more important or larger than a difference significant at the .05 level. Referring to a difference as "highly significant ($p < .001$)" is faulty reasoning and an inappropriate use of the decision rule. The significance level depends on sample size. Even if sample sizes were equal the p-values describing the decision rule do not provide an index of the *actual size* of the difference or effect.

A better way of presenting the results is to calculate the *magnitude* of the effect, known as the *effect size* interpreted relative to the spread of scores. For example, the difference between two means may be divided by a pooled estimate of the standard deviation. Effect sizes give an indication of the response to the question “How big a difference is it?”.

Table 4 shows that an effect size of 0.00 to 0.14 is **very small**. (The two group means are almost the same or the same; there is almost total overlap between the two score distributions for the groups.) An effect size of 0.15 to 0.44 is **small**. (The two group means differ; the overlap between the two distributions is considerable.) An effect size of 0.45 to 0.74 is **medium**. (The two group means differ; the overlap between the two distributions is moderate.) An effect size of 0.55 or more is **large**. (The two group means differ; the overlap between the two distributions is small.)

Table 4 Descriptors for magnitudes of effect sizes (after Cohen, 1969, p.23) and assigned ranges (Izard, 2004)

Effect Size Magnitude	**Cohen’s Descriptor and Cohen’s Example**	Assigned Range
< 0.2	Very small*	0.00 to 0.14
0.2	Small difference between the heights of 15 year old and 16 year old girls in the US	0.15 to 0.44
0.5	Medium (‘large enough to be visible to the naked eye’) difference between the heights of 14 year old and 18 year old girls	0.45 to 0.74
0.8	Large (‘grossly perceptible and therefore large’) difference between the heights of 13 year old and 18 year old girls or the difference in IQ between holders of the Ph.D. degree and ‘typical college freshmen’	0.75 or more

* Note that “very small” is a descriptor devised by Izard for magnitudes less than “small”

Tables of results should include effect sizes (Cohen, 1969). The appropriate statistics for educational research are point estimates of effect sizes and confidence intervals around these point estimates (Jones, 1955; Kish, 1959; Rozeboom, 1960; Carver, 1978; Hunter, 1979; Oakes, 1986; and Schmidt, 1994). Substantive significance is more useful than statistical significance. For example, when a new teaching method is compared with another method, the magnitude of the difference between the methods may not be large enough to be worth the expense of changing methods.

Possible explanations for these concerns about conference papers remaining current for more than a decade include

- cultural lag (because researchers/supervisors are not up to date in their reading on these issues – note the dates of the references in the paragraph above),
- failure to choose referees with expertise in methodology and related topics,
- a lack of quality control over the refereeing process as a consequence of the anonymity of referees,
- restricting referee comments to the person(s) offering the paper (rather than making these comments available to readers of the paper),
- failure to include such issues in research methodology courses requiring participation by candidates for Masters and Doctoral research,
- an unjustified assumption that a pass in a research methodology unit means that a student is competent to use multiple methodologies,
- choice of a single methodology by post-graduate researchers before choosing a topic (rather than choosing a topic and the adopting relevant methodologies),
- supervisors being pressed to take on candidates with topics outside the supervisor's expertise,
- blind adoption of software-related decisions without considering whether the study group is a population, a representative simple random sample, a representative complex sample, or intact classroom groups,
- an expectation that software will somehow overcome inadequacies of experimental or quasi-experimental design, instrumentation (whether test, rating scale, checklist, or questionnaire) and choice of analysis strategies, and
- failure of conference committees to advise adequate academic requirements for research papers.

Problems perceived in **research journal papers** (and research theses, for that matter) submitted for peer review (or external assessment in the case of theses) over a similar period also include many such issues. Possible explanations for these concerns remaining current for more than a decade include many of those listed above for conference papers, together with

- non-transparent processes of choosing referees,
- a lack of accountability of referees,
- an unjustified assumption that the number of rejected papers is automatically an indicator of quality, and
- delays in the publication process (peer review, acceptance/invitation to re-submit/rejection/loss of manuscripts, editing, permissions for quotations and artwork, proof-reading, and production) that could mean papers are not available for years.

Table 1 Problems in proposed papers submitted for peer review refereeing for AARE conferences (Source of information: IZA06033)

Problem	Examples
Flaws in methodology: Inadequate research design	• Lack of concise and precise detail about the research intentions • Failure to specify details of differential treatments • Failure to indicate whether participants were assigned at random or as intact groups • Lack of detail about comparisons • Failure to provide evidence that groups were comparable at the start of the study • Failure to consider threats to validity inherent in the research design
Flaws in methodology: Assessment strategies not validated prior to use	• Using instruments from other research without checking that they are valid in this context • Constructing new instruments and gathering data without validation • Not considering whether meaningful differences can be detected for this time span • Ignoring ceiling and floor effects of using narrow-range instruments • Sparse data reports based on minimal evidence
Flaws in methodology: Inappropriate measures of learning	• Must measure on more than one occasion to demonstrate learning (see also IZA04877) • Must have at least two relevant *comparable* measures (see also WAT04867) • Instruments need to cover what is known now as well as what is intended to be learned
Flaws in interpretation: Claiming that learning has occurred in Spite of the evidence collected	• Inferring learning from a single achievement measure • Inferring learning from two separate un-calibrated tests • Very small differences judged as significant
Flaws in interpretation: Claiming that absence of learning is a consequence of a treatment	• Ignoring ceiling and floor effects • Relying on a single achievement measure • Ignoring non-assessed learning

Table 1 Problems in proposed papers submitted for peer review refereeing for AARE conferences (Source of information: IZA06033) (continued)

Problem	Examples
Flaws in interpretation: Inadequate consideration of competing plausible explanations for treatment effects	• lack of controls (whether prior achievement or control group comparisons) • failure to consider alternatives
Inappropriate application of statistical significance Faulty reasoning	• complex samples assumed to be simple random samples when judging statistical significance • *not* for populations (since one does not need to infer a population value if it is known) • interpreting failure to reject null hypothesis as "no difference"
Failure to indicate the magnitude of effects Faulty reasoning	• substantive significance confused with statistical significance • claiming big effects for trivial differences • large effects may not be reported

For example, if referees are chosen from a restricted list, the group of referees is not representative of the peer group of researchers. Small groups of referees allow manipulation of the review process and threaten the validity of the judgments. If referees are accountable to a small group from one or a limited number of universities, the tendency is to accept only papers acceptable to that university group. This has an impact on the viability of competitors regardless of the quality of their output. This is contrary to ethical use of peer refereeing.

The assumption that the number of rejected papers is an indicator of quality rests on a further assumption that all referees are consistent, valid, and equally stringent judges of quality for all their refereeing duties and that the criterion for an accepted paper is stable regardless of the number of papers submitted for assessment. If the proportion rejected is consistent regardless of the quality of the submitted papers (as judged independently), both assumptions are invalid. Reporting the number of rejected papers without a criterion-referenced interpretation of quality gives no information about the legitimacy of the cut-off between acceptable and unacceptable. As such, there is no justification for the decisions, nor can there be any indicator of the capacity of the referees to distinguish between acceptable and unacceptable papers. If we cannot demonstrate that referees can make such a distinction consistently over time, the refereeing process is arbitrary, indefensible and unethical.

Delay in the publication process is a potential contributor to variation in quality: over an extended period it is difficult for referees to be objective in applying standards in a consistent manner over time.

Referees face difficulties in judging the quality of proposed papers for AARE conferences as shown in Table 2. Some of these difficulties arise from restrictions of access to journal papers that are a consequence of limited economic resources in universities and other research agencies.

Table 2 Problems faced by reviewers of proposed papers submitted for peer review refereeing for AARE conferences (Source of information: IZA06033)

Problem	Implications
Restricted access to references quoted in the paper: Example 1 - Access denied. You may not have a subscription to this journal. If this is the case, you can view and download the PDF of this paper using the pay per view option. Online access is restricted to institutional subscribers. This is done by restricting the IP (Internet) addresses that are able to access the full journal Example 2 - You are trying to access material included in JSTOR, an online journal archive made available to researchers through participating libraries. Unfortunately, you do not have access to JSTOR from your current location	• Should papers with 'close' references be rejected automatically? (Few universities and research agencies can afford all journals) • Cannot judge quality from an abstract • Who pays for inter-library loan or pay per view? • Should those submitting a paper for review be obliged to attach copies of relevant parts of referenced papers? (Would this breach the original copyright in some instances?)
Heavy work load imposed on referees Example - 600 papers would require a minimum of 1800 reviewers, more than the entire membership of AARE	• Researchers unwilling to referee papers, particularly when not acknowledged • Unwillingness to take on unpaid work • Difficulty in meeting tight deadlines, tendency to take short cuts and avoid detailed comments • Difficulty in getting sufficient qualified referees • Those who do accept have to cover for the unwilling • Pressure on managers trying to keep to deadlines

Other problems are a consequence of excessive workload now expected in academia. Refereeing is regarded as unpaid and un-acknowledged work, requiring sustained effort and is subject to tight deadlines. All of these problems apply to journal referees as well.

Editors and conference organisers have different but related problems. When proposed papers are received they need to find a match between the expertise of their list of referees and the content of the submitted paper. This is made more difficult by

ethical considerations such as conflicts of interest, and issues that are the academic equivalent of "Commercial – In confidence". Obtaining sufficient competent and suitably qualified and experienced referees that meet the ethical considerations is a difficult task given the size of the research community and the predilection of university funding agencies to restrict the number of research centres with interests in similar areas.

Once referees have been identified for a particular paper, the next task is seek their agreement to do the task. It may take 10 invitations to obtain 3 referees prepared to act and who actually carry out the task for a conference paper (JEF07611, p.5). I do not have recent evidence about the corresponding difficulty facing editors but earlier experience suggests that the difficulties are comparable. Failure to respond to invitations to referee or to meet deadlines places additional pressure on conference managers/journal editors because their deadlines still should be met to keep faith with the membership and subscribers.

This is a "people" problem faced by existing procedures for AARE conferences as well as the proposals for the 2008 AARE conference. Adopting different software will **not** solve this problem as software cannot make researchers more compliant when a task is unpaid, stressful because of a deadline, and the professional work is not acknowledged. Making deadlines tighter will only exacerbate the difficulties and/or lead to a reduction in the quality of the refereeing process in order to meet deadlines. (One cannot delay a conference depending on refereed papers if the refereeing cannot be carried out on time. Some journals are notorious for being late but whether this is due to referee tardiness or editorial management is a matter for speculation.)

This difficulty in obtaining suitable referees providing timely and informative reports is only an initial step in making sound decisions. Unless referees make valid and consistent judgments distinguishing between different levels of quality, the information gathered will not be useful in choosing quality papers for publication. Although we might not expect experts to agree fully on all aspects in their review there must be some general independent consensus between their respective judgments. If referees are inconsistent with each other, this implies that they are assessing to different standards or using different criteria (implicit or explicit). Is this evidence of bias? How are such referees to be held accountable? Should referees be required to provide constructive professional comments? If they generally agree but differ in stringency, which views should prevail? Table 3 shows a range of such problems for editors and conference organisers, and lists implications of these problems.

Some journals seek accountability by requiring them to be identified in the same way as the author. The submitted paper, the referee comments and revised paper is available for all to see. The reader may judge the quality of both the author's work and the referee comments. I return to this issue later in this paper.

Are current and proposed procedures adequate?

By current procedures I refer to the refereeing system existing from 1999 to 2007. This system has evolved to meet new demands over time.

Table 3 Problems about referees faced by conference committees or journal editors (Source of conference information: IZA06033)

Problem	Implications
Matching referee expertise with content of proposed papers	• Need a comprehensive and current list of competent referees • Assumes key words are commonly accepted and have same meaning nationally/world wide
Conflicts of interest (supervisor, colleague, family member, chairs of promotion committees, competitor for research grants)	• Rule out same institution (and State?) • Rule out family • Rule out staff at competing unis?
Obtaining sufficient qualified referees	• Knowledge of each field or have a network of contacts with that knowledge
Ensuring that referees have access to relevant research literature	• Need to avoid potential conflicts of interest (those most likely to have access to the same literature are competitors, former colleagues or family members) • Other universities may not have access to referenced journals
Ensuring that referees make valid and consistent judgments, distinguishing between different quality levels	• Can this be assumed? • If not, how might it be checked? • What happens if inconsistencies are found? • Inconsistencies among reviewers are ignored: not a "transparent" process (bias is not detected)
Ensuring that referees are equally stringent/lenient	• Research shows that this is most unlikely; what adjustments should be made?
Guaranteeing that referees provide constructive professional comments	• Should comments be published with the paper? • Should each referee be identified by comment?
Sharing referee expertise beyond individual authors - Only the author has a chance to address a problem identified by reviewers: other authors lose this opportunity to learn (they do not see the comments)	• Current research culture is more than 7 years out of date: errors are repeated year after year (Is failure to address this ethical?) • Should a condensed version be available to future authors?

Its first strength is the large number of referees grouped according to specialist expertise and readily identified through use of widely-accepted key words in the Australian Thesaurus of Descriptors, used in the Australian Education Index (and

consistent with similar descriptors used world wide) and AARE SIG titles. The second strength is the capacity to arrange the refereeing process in concert with the timetable for the collection of conference registration fees in time to confirm the size of each conference (and therefore the number and size of rooms required) and to pay contracted deposits on time without affecting AARE finances. Its third strength is that it can be managed by a non-academic without giving control to representatives of any one university. That has consequences for lower costs too (because no high academic salaries and on-costs have to be met).

Its weaknesses are the dependence of the conference system on conference committee reviewer and referee goodwill, the brevity of the 4-item evaluation scale developed from the referee procedures of the Australian Educational Researcher (see JEF07611, Appendix 6) and the lack of consideration of the quality of the referees and their accountability. Similar approaches for journals also depend on referee goodwill and may or may not exercise quality control over referees.

The description of the proposed procedures for 2008 has only been available for a short time in the brochure on the AARE web site (flyer1.pdf obtained from http://www.aare.edu.au/conf2008/index.htm). It is not clear, on the information I have, what new software will be used, who will control the procedures, what cost this will entail for software, management and (possibly) academic supervision, or how referees will be selected and encouraged to work with a shorter time frame later in the academic year. Even though I was the AARE-NZARE conference treasurer at the Melbourne Convention Centre in 1999, it is also not clear to me how the refereeing process has been arranged to match the timetable for the collection of conference registration fees in time to confirm the number of paid participants and to pay required presentation, meal, and venue deposits (with reasonable accuracy) on time according to legal and binding contracts without affecting AARE finances and exceeding conference budgets by thousands of dollars.

As I said last year at the Adelaide AARE conference, it seems ironic that a system of peer review can place so much emphasis on judging the quality of proposed papers without considering the quality of the judges *at the time they make the judgments*. Given that AARE is a professional body of researchers one might expect evolution of the 1999-2007 system to address quality issues with respect to referees. I am surprised that a new system appears to have been introduced by persons assumed to have research expertise without what I see as essential research and validation, followed by cost-benefit analysis.

What improvements are necessary?

It is my belief that capricious changes based on opinion rather than sound research evidence are the antithesis of what might be expected of a body of experienced researchers. I recognise that improvements are required, but believe they should take prior experience into account. I also believe that the claims of competing systems should be subject to rigorous analysis and investigation before a new system is adopted. Some examples follow.

MyReview

At last year's AARE conference in Adelaide I reviewed the MYREVIEW System (Version 1.9.3) by Philippe Rigaux of Laboratoire de Recherche en Informatique Université Paris-Sud Orsay, France <rigaux@lri.fr> dated July 26, 2006. I noted that

- Reviewers were identified by individual email addresses & a list of mutually exclusive research topics (Rigaux, 2006, p. 9)
- MYREVIEW tries to determine conflicts based on authors and reviewers names and affiliations, and
- The program available at that time was limited for large datasets (around 200 papers) (Rigaux, 2006, pp. 26-27). I noted that AARE has around 3 to 4 times that number in a good year.

I also note that the MYREVIEW Manual says, "You must let your PC members browse the list of papers and refine the automatic preferences". In effect this means every *reviewer* has to read the *abstract* of every paper before *papers* are assigned to *reviewers* in order to determine whether they wish to declare a conflict or an indication that they do not wish to review that paper. (In these days of multiple e-mail addresses with gmail, hotmail and yahoo, the task of avoiding conflict of interest in referee invitations is not easy to address.) The MYREVIEW weighting algorithm adds the weights offered by *reviewers* for each paper. A paper is assigned to the 3 reviewers who gave the highest number. If they do not know the name of an author and the university, how can they declare a conflict of interest? If they do know the name, how can it be a blind review?

A further concern was that the *administrator* can alter the weights for certain *reviewers* manually so that they are never offered papers to review even though they have been asked and have agreed. (See section 5.3 of the latest Manual) [The *administrator* will still be able to claim this is peer review because all were invited but DEST or the New Zealand equivalent may not be aware that the file may be modified to reduce the pool of *reviewers*.]

I also suggested that if the procedure was to be modified, some consideration should be given to the processes adopted by other on-line journals, such as *Virtual Physics* (Swinburne) and the *Journal of Interactive Media in Education* (Open University). Open source software for online journal and conference systems such as that from the Canadian Public Knowledge Project (see http://pkp.sfu.ca/) should also be investigated.

Virtual Physics

This journal is an on-line forum for virtual meetings of scientists and students involved in a research activity on contemporary physics. The editors are from Belgium, Sweden, Australia, Canada (Webex), Poland, and USA. (See http://www.swin.edu.au/chem/complex/vp/vp12/vp12.html) It includes a lively debate on "Research and Funds" with a September, 1996 contribution on "Peer Review" by Alexander Berezin from Canada.

Journal of Interactive Media in Education

This journal (based at the Open University in United Kingdom) uses a sequence of review stages

1. private, open peer review by three named referees acknowledged for their contribution to a review, with interaction between author and referees;
2. public, open peer review with readers able to review the debate and contribute at this preprint stage; and
3. final publication freely accessible and linked to further debate and research). Figure 1 shows this sequence of events in diagrammatic form

(see also http://www-jime.open.ac.uk/doc/reviewer-guide.html).

According to the proponents, "In conventional journals, the point of publication is the beginning of scholarly debate. JIME brings this point forward by making submitted preprints accessible, but of course *continues to support discussion about the revised, published article.* In addition, the most interesting review comments/exchanges are published with the final version, providing readers with insight into the issues that arose during review, and enabling them to build on those discussions. … The final publication will be freely accessible on the JIME site. "

This model depends upon the willingness of both authors and referees to engage in professional debate, initially in private, and then in public, moderated by the editor.

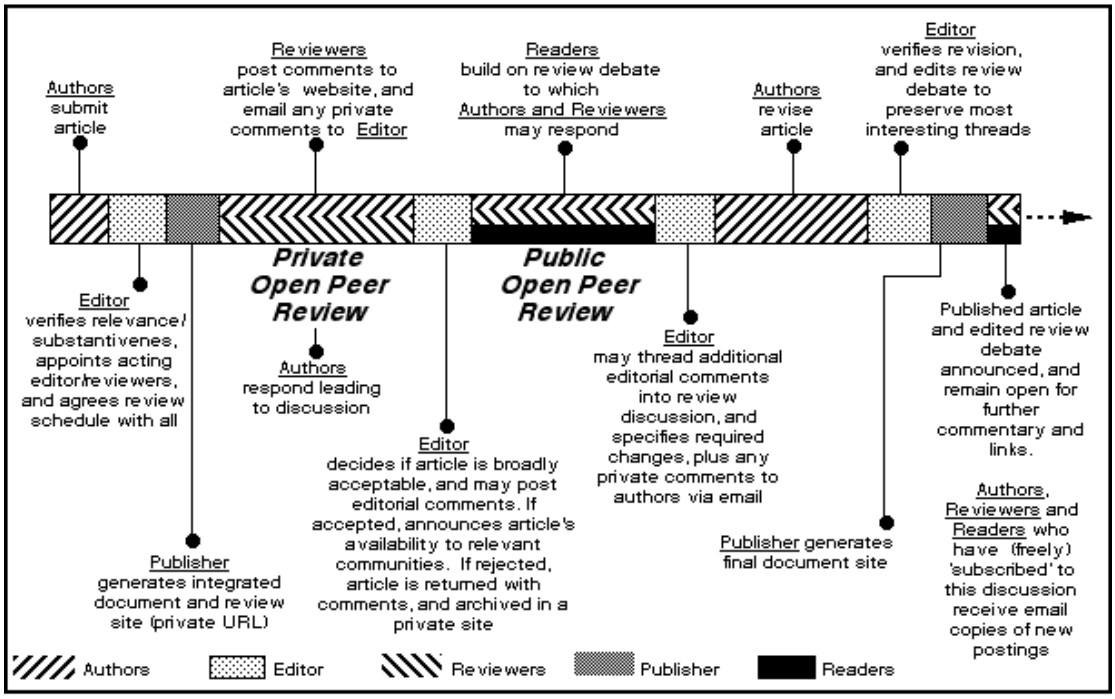

Figure 1. Lifecycle of a JIME submission

It is claimed that, "The philosophy behind this model is that perceived risks of this sort will be outweighed on the one hand, by the benefit to authors of quicker and more extensive feedback, and on the other, the increased opportunity for peers working in the field to critique and shape a submission before it is published. Authors have

reported that they have greatly valued the discussions that have emerged. Ultimately, we hope that these forces can converge to create higher quality contributions."

Public Knowledge Project

This Canadian federally-funded research initiative seeking "to improve the scholarly and public quality of academic research" is based at the University of British Columbia and Simon Fraser University. The Public Knowledge Project (PKP) has developed free, open source software for both conferences (see http://pkp.sfu.ca/?q=ocs) and journals (see http://pkp.sfu.ca/?q=ojs). The conference software is a web publishing tool on the PKP server still under development: session scheduling, reviewer form, and associated help files are described as upcoming features. By way of contrast, the journal software is installed and controlled locally. Since this can be downloaded as open source software, a local site may make code changes under a general public license. But like other open source systems for journal management and publishing (many listed on the PKP web site), this large system comes with no warranty or guarantee of support. The FAQ section on the open journal system web site makes it clear that a single journal installation has clear advantages over a multiple journal installation, and that technical expertise is required to adapt the software to suit the server, and the other software modules on that server. If proposing to use such a system, it would foolhardy to adopt it before trials have been conducted. It makes sense to have professional expertise implement such software on a single AARE site so that all necessary automated backup procedures, database drivers and third-party software can be loaded and maintained, without the additional expense of transferring the system to different servers with every change of editor or editor's change of institution.

Conclusion

The AARE provides a mechanism for members to share knowledge, seek feedback from colleagues, and to acknowledge those papers that have met set criteria for scholarship. This mechanism applies to both conference papers and journal papers although the details may vary. The feedback from referees when papers have been submitted could be considered part of the benefits of membership, although there is no requirement to be a member when submitting a paper for a conference or the *Australian Educational Researcher*. The question of whether AARE members should fund the provision of referee services for the conferences and/or the *Australian Educational Researcher* needs consideration. At present, since rejected papers may lead to the non-member potential presenter failing to attend the conference, a moderate fee is sought to cover administrative costs of the refereeing process. This avoids members subsidising non-members to some extent although the web site places accepted and delivered papers for all to see if the authors lodge the papers for this purpose.

When evaluating the quality of reviews it is difficult to escape considering the choice of referees, the conditions that referees are asked to meet, and the quality of the judges providing those reviews. The evaluation component of the AARE reviewing form is the equivalent of a 4-item partial credit rating scale. This is not adequate for valid ratings of quality conference papers or journal papers and should be expanded to

achieve valid evaluations. Further, the quality of the judge agreement should be investigated for each paper so that one may be assured that ratings vary sufficiently to be independent assessments, but show a degree of consistency to be sure that the judges are making comparable judgments.

References

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press

Hunter, J. E. (1979, September). Cumulating results across studies: A critique of factor analysis, canonical correlation, MANOVA, and statistical significance testing. Invited address presented at the 86th Annual Convention of the American Psychological Association. New York, NY.

Izard, J.F. (2004). Best practice in assessment for learning. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on *Redefining the roles of educational assessment*, March 8-12, 2004, Nadi, Fiji: South Pacific Board for Educational Assessment.

Jeffery, P.L., (2007). AARE's Conference full papers academic peer refereeing processes. Paper presented at the AARE Conference in Fremantle, Nov. 2007. (http://www.aare.edu.au [search code JEF07611]). Melbourne, Vic.: Australian Association for Research in Education.

Jones, L. V. (1955). Statistics and research design. *Annual Review of Psychology*, *6*, 405-430. Stanford, CA: Annual Reviews, Inc.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review*, *24*, 328-338.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: John Wiley.

Ross, K. N. (1975). Searching for uncertainty: an empirical investigation of sampling errors in educational survey research. Unpublished M.Ed. thesis, University of Melbourne.

Ross, K. N. (2005). Sample Design for Educational Survey Research: Module 3. Paris: International Institute for Educational Planning (UNESCO). http://www.sacmeq.org/training.htm or http://www.unesco.org/iiep .

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Schmidt, F.L. Presidential Address to the Division of Evaluation, Measurement and Statistics (Division 5 of the American Psychological Association) at the 102nd Annual Convention of the American Psychological Association, August 13, 1994, Los Angeles, CA.

Thompson, B. (1999) 'Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.' Invited address presented at the annual meeting of the American Educational Research Association, Montreal. [http://acs.tamu.edu/~bbt6147/aeraad99.htm or www.coe.tamu.edu/~bthompson/aeraad99.htm]