Is the MERGA Conference Refereeing Process Fair?

John Truran
Private Practice, Adelaide
<truranjk@camtech.net.au>

Kathleen Truran
University of South Australia
<Kath.Truran@unisa.edu.au>

This paper examines the extent to which MERGA's 1999 Conference refereeing procedure was fair to the authors. The short answer to this is "yes". But while conceding that the realities of preparing a conference mean that fairness must be balanced against several constraints, the paper shows how some weaknesses may be ameliorated by detailed planning and a little more editorial authority.

MERGA's four-yearly reviews of Australasian mathematics education research (e.g., Sullivan, Owens, & Atweh, 1996) provide good summaries of activities by MERGA members and others, much of which has been presented in some form at MERGA Conferences. But the process by which this research has reached publication status has received little formal attention even though peer refereeing is seen as a *sine qua non* of academic endeavour.

MERGA Conference Proceedings have been refereed since 1993, with acceptance requiring a majority decision from three people described in the annual calls for contributions as "senior researchers". The number of referees used has gradually increased, without any corresponding increase in the number of accepted papers. The 1999 Conference used nearly twice as many referees as in 1998, and not all of them were "senior". This diversification was required by increasing work and publication time pressures, and was welcomed by some members. Here we use findings from confidential data available to us as editors of the 1999 Conference Proceedings (J. & K. Truran, 1999) to assess the balance between widening the refereeing base, maintaining the quality of papers, and fairness to authors of rejected papers. We report in a form which illustrates finer points of the process while preserving anonymity.

The MERGA Conference Refereeing Process

Each year the process is administered by host city editors who have a high degree of autonomy while still subject to MERGA and Conference structures. Their academic role involves matching submitted papers with appropriate referees, checking that referees' reports are appropriate, and seeking advice in borderline cases before making final decisions. Since a wide range of research approaches and paradigms is encouraged, this task is not always easy.

Refereeing is "double blind", but MERGA is small and many authors will be recognisable. There is an "Early Bird" (EB) facility, principally for new researchers, but available to all, allowing referees to recommend revision and resubmission of a paper. But for the remainder, referees may give only "accept" or "reject" judgements. Referees are generally sent no more than three papers, but their work is a labour of love, relying heavily on members' substantial goodwill. They are provided with the criteria authors were expected to address and a pro forma listing seven aspects to address before making a judgement.

Because of tight time-lines, we sent all papers to three referees simultaneously—a break with the previous practice of using a third referee only for hung judgements. This proved to be wise. The experimental probability that any two out of three referees would provide identical decisions was about 0·45. Some differences were academic ones, but others arose from failure to return or acknowledge the receipt of papers, unwillingness to referee a specific paper on ethical grounds, etc. Given that the 0·45 includes some reports which were returned very late, using only two initial referees per paper would have required a time-consuming second mailing for about two-thirds of the papers.

Choosing a Pool of Referees

We estimated that for a projected submission of 100 papers we would need at least 100 referees—a doubling of a refereeing pool which had never been larger than 51 in the past. We had attended most MERGA Conferences since 1992 and knew most long-standing members by sight, but often knew little about their research interests. So we compiled a data base of all current members, all authors of refereed papers to MERGA Conferences between 1993 and 1998, and all referees used (excluding 1997 when all referees were from New Zealand and no list was provided). The MERGA Membership Directory (White, 1998) with its electronic update provided members' reported "major areas of interest" (usually three or less) and "current research" and also lists of who was interested in each research field.

After discussion with the local Programme Committee we decided that the basic criterion for the refereeing pool should be two refereed papers already accepted for MERGA Conferences. This criterion satisfied the Department of Education, Technology and Youth Affairs requirements operating at the time, and was more rigorous by one paper than that used for the international Psychology of Mathematics Education (PME) Conferences.

Of our basic list of 134 referees, 49 had not refereed before. We added another eight members who were distinguished researchers but had been little involved in MERGA Conferences and four distinguished unfinancial members. We did not include a further 32 eligible, but unfinancial, researchers, although five of these did submit papers in 1999. Thirteen former referees did not have sufficient published papers to be included in our list—indeed, eight of these had never presented a refereed paper at a MERGA Conference. So our refereeing pool of about one-half of MERGA's 300 members was constructed in a way which seemed fair to authors because it was based on transparent criteria, and fairer to members, because it made it easier for newer members who had some proven research experience to be recognised as ready for refereeing experience.

Indeed, after we had written to members on our list, several new referees said how pleased they were to be invited, and a few members sent favourable comments about our general approach to selecting referees. Some of those selected were unavailable in 1999, and a small number of experienced members with many MERGA Conference publications declined to be involved, but we were confident that we would have sufficient referees. In practice, other difficulties did arise. Some were unavoidable—study leave, sickness, unexpected other commitments, etc; others arose from human frailty—long-delayed or cursory reports—or electronic problems. Some reports were never returned or acknowledged. It proved necessary to use four South Australians who did not fit the standard criteria, but who had had relevant experience.

We did not find that inexperienced referees necessarily produced poor reviews. Many did not. In the important case of ultimately rejected papers there were only eight new referees concerned with fourteen papers, and only two papers were read by more than one new referee. Some of these had had extensive experience in other areas and, to the best of our knowledge, none of the others was naïve. In any case the vast majority of negative reviews came from the more experienced referees. Using many new referees did not affected the ostensible fairness of the process at its most critical point.

This section has summarised how we doubled the size of the refereeing pool in a systematic way which ensured that all referees had had at least minimal adequate experience for the task. We now need to assess the extent to which this change provided fair judgements for authors.

Questionnaires

Two questionnaires were prepared. Responses will be used at suitable places in the analysis below. The first was given to all conference participants and attracted 18 responses. It included the following questions:

- If you were a referee did this system mean that you reviewed papers that were appropriate to your interests and specialties?
- As an author do you think your paper was reviewed by a person confidant in your specialty?
- If referees were not unanimous in the judgement of your paper how did you respond?

The second was sent to thirteen authors of rejected papers and asked if they considered their referees to have shown a sound grasp of the field and to have made reasonable points. Where the referees were not unanimous we also asked whether the positive review was seen as more valid than the negative ones. We received seven responses, including four from senior MERGA members. ("Senior" from now on refers to a person who has either held a leading academic role in MERGA or has a rank of at least Associate Professor.)

The Fairness of Judgements on Papers Which Were Rejected

In this paper we focus on the fairness of the process to those whose papers were rejected since this is the most overt form of potential error. But acceptance of poor papers is also unfair, certainly to the authors, and also to those who attend the presentations. Indeed, we had received comments about papers which were believed to have slipped through the net in previous conferences; unfortunately, space does not allow further discussion of this point here.

Some rejected authors were understandably upset, and some very quick to withdraw from the Conference. One believed the rejection was the result of paradigm warfare; another sent an unsolicited short list of suspected referees! It was totally inaccurate. Several believed that some of their referees did not understand their work or their paradigm very well, particularly if they were not working in traditional moulds (either quantitative or qualitative).

This important claim was also made by some of the accepted authors. It is difficult to examine and hard to rectify if valid. As editors, we could not have pre-read every paper: many would have been too far beyond our expertise for us to have anticipated difficulties anyway. But some criteria are available.

Acceptance Rates

One simple measure of fairness is acceptance rates, which could show if there has been a substantial, not easily explained, change from previous years.

In 1999 there were 17 EB and 70 standard submissions. Eight EB papers were accepted at once, and the others resubmitted and subsequently accepted, apart from one withdrawn for personal reasons. Four EB and seven standard papers were entered for the Practical Implications Award which was judged separately by a panel of five who also accepted or rejected the papers. Our main editorial task dealt with 80 papers, sent to 96 referees, of whom 90 returned reviews.

Of the 86 final proposals received, 70 were accepted, and 16 recommended for presentation as Short Communications. Seven of these papers, from six authors, were withdrawn, in four cases accompanied by the authors' withdrawal from the Conference. So the acceptance rate for 1999 of 80 percent is similar to that of 84 percent for 1996 (Clarkson, 1996). Data for other conferences are not available, but Clarkson suggested that the rate seemed to be reasonable to him, although he did receive some comments that it was unreasonably high.

So the acceptance rates suggest no large divergence from a general practice which had been seen as reasonably effective. The rejection of at least five papers by senior authors was not surprising, because a similar phenomenon had occurred in 1997 (Biddulph, co-editor, pers. comm.), and Clarkson (1996) had noted that in some jointly-authored papers "the senior partner had [clearly] given little input into the final development of the paper". Three of our rejected papers may have fitted this situation.

Quality of the Referees

Peer review relies on consensus among experienced researchers for its validity. No editor can be sufficiently knowledgeable to assess fairly all papers and judgements. A high quality of referee is essential. So we shall examine the quality of our review process in several ways as basis for a general conclusion.

Comparison with Other Areas of Mathematics Education

Of the 90 referees whose reports were returned, 28 were either members of the Editorial Board, or in 1999 had refereed for MERGA's *Mathematics Education Research Journal* (MERJ, 11, 2, 156). Eight (out of sixteen Australasian referees) were used by the American *Journal for Research in Mathematics Education* in 1999 (JRME, 30, 5, 598–599), and one (out of five) by the European *Educational Studies in Mathematics* (ESM, 40, 3, v). Referees used by these journals but whom we did not use were either unfinancial, unavailable, not within our criteria, or just not called upon.

Two points arise. Our refereeing base, as in all recent MERGA Conferences, used far more members than do several leading journals. This says little about quality, because the selection of referees reflects editorial knowledge and experience. For example, the small proportion of Australians used by JRME does not reflect their large international contribution to mathematics education.

More importantly, many senior MERGA members were "unavailable". Their reasons were many, some obviously reasonable, some less so. For conference papers there is a need for the greatest expertise possible. Judgements are more absolute and immediate than those on journal papers. Every negative decision affects a university's research budget by about $1000, may affect the Conference's budget, and will have a significant irrevocable personal impact on authors. We had some weak referees, and our widened base would have benefited from the presence of more very experienced researchers.

The Judgements of Specialist Referees

Some evidence for this claim is available because we tried to use referees with specialist skills who were outside (but only just outside) the normal MERGA circle. Three well qualified people co-operated, refereeing six papers from five authors. They rejected all six. In three cases all referees were agreed, in two others the paper was rejected and in one case the paper was accepted.

This small sample cannot provide conclusive evidence, but does suggest that papers sent to an outside expert were more likely to meet a more demanding standard than those sent to people, however well read, with only a general interest in the topic. The fact that in the three non-unanimous cases the "expert" opinion was that the *methodology* of the research was inadequate lends support to this claim because such technical detail is unlikely to available to the general reader. Indeed, two of the methodologically poor papers received perfunctory positive reviews from a long-standing MERGA member with a professed interest in the field, but no relevant published papers, at least at MERGA conferences. The contrast in quality

found here provides a good case for some editorial discretion over accepting minimalist reviews (our shortest was one word), especially when they are negative. It also suggests that regular infusion of new blood is likely to lead to an increase in standards over time.

Matching Papers with Referees

Some of these difficulties could be avoided by making a good match between authors' content and referees' expertise. Most respondents felt that the reviewers were competent to review their paper, although one wanted more detailed feedback to be sure that the referee had actually read the paper, and one felt the referee didn't know enough. Equally, most referees were willing to review papers not within their expertise, no doubt aware of the difficulties we were encountering in effecting good matches between our data base of interests and authors' description of their papers in terms of MERGA's categories. This was made more difficult because some information for both authors and referees was not available, forcing us to make some classifications ourselves. as well as making some adjustments because the classifications had changed in the period just before the 1999 Conference.

Assessing the quality of matching is not easy. One-one correspondence will not do, because a paper may not be adequately classified. Thus a paper classified as "Geometry" but with an assessment perspective might lead us to choose one referee with expertise in assessment but not in geometry. Sometimes we made use of personal knowledge not recorded on the database, particularly to ensure that good researchers were fully utilised. Of the fourteen rejected papers we handled, copies were sent to 47 referees. Five went to outside experts, as discussed above, 23 to good matches in the classification system, and seven to mature researchers with a broad interest. Seven went to referees for reasons which are not apparent from the data base ten months later—possibly creative desperation, and five were not returned by referees, requiring a second, perhaps less good, choice of referee. This was the best we could achieve: the work was done thoughtfully and never rushed. But some check was available on poor matches by examining the quality of responses, as discussed below.

One additional problem arose because many researchers work in teams which submit several papers. We decided to send papers from one team to no more than one member of any other, and preferred to send papers to three different states or at least three different universities. These constraints caused special difficulties for smaller research fields. In one field, for example, there were five appropriate referees, two pairs of whom formed research teams. We distributed the two papers from a third research team as fairly as we could, but it happened that for one of these papers one referee with expertise did not respond, another rejected the paper, and the third, with no specific expertise, accepted it. The ring-in, last-minute, local, fourth referee had no specific expertise and was quite aware that a delicate line-ball judgement was being called for.

Proven Experience of the Referees

Did members' expressions of interest match their published papers and proven experience? A full answer would require examination of each referee's *corpus,* which has not proved possible. But a random sample of ten referees was taken to compare expressed interests with refereed papers published at MERGA conferences. This small survey is sufficient to indicate the existence of all five theoretically possible cases. One person had interests and publications which matched exactly, and two had no overlap between them at all. Three had interests which were wider than their published papers, and one had papers not related to their declared interests. Three had both of these: papers outside their interests and interests outside their papers.

Such diversity is not surprising. Our own listed research interests cover three where we have published papers, another where one of us has a major piece of work in preparation, and two where one of us is currently teaching students and is therefore familiar with the literature. No doubt others have expressed interests for similar reasons. But to find such diversity in a small sample is strong evidence that a process of selecting referees based on their proclaimed interests does not ensure that they have *proven* research expertise in those fields, and also fails to pick up expertise which has been proven.

The data presented here show that some of the real difficulties encountered in making good refereeing matches might be overcome by a more detailed data base, both of skills and of past reliability. Weaknesses which remain may be partially addressed by more attention to the quality of the reviews.

Assessing the Quality of Review

We discuss the general efficacy of the reviewing process elsewhere (K. & J. Truran, 2000); here we discuss some hard cases. We have mentioned how hard it is to assess work in unfamiliar fields. Nevertheless, for us good reviews usually conformed with the published criteria, showed awareness of other work in the field, summarised strengths and weaknesses, and made an holistic summary arguing a case for the judgement given. The best ones also suggested improvements and ideas for further work. The poor one contained errors of fact, made unreasonable demands on a short paper, were perfunctory, or offered little of value to the author. Frequently they were also returned tardily.

As a check, one of us read all reviews to check for glaring inconsistencies, and both of us read all reviews of potentially rejected papers. We found two situations which we thought needed attention. The first was when a review did not seem to be factually accurate. For example, one negative review stated that a theoretical paper was not "sufficiently concerned with mathematics education" and "did not offer a significant review of a body of literature [on] mathematics learning". This strong claim matched neither our knowledge of the author nor what we read in the paper, so we sought a fourth opinion on the accuracy of the claim and the acceptability of the paper, because the other referees were divided. Strictly, we went beyond our remit, but we considered that we had authority, after suitable consultation, to over-rule a report on a matter of fact. In our opinion this produced a much fairer decision.

Secondly, we found a minority of reviews which were brief, trivial, rushed, or more than one of these. One example has been mentioned above. We wanted to put these to one side, especially the negative judgements, and to seek further opinions, but time did not allow this. It was a consolation to find that only one of the negative reviews of the rejected papers was close to being unacceptable on these grounds, but this did at least cite clear, testable reasons for rejection, albeit very briefly.

We are not of the opinion that editors should wantonly put any reviews aside. The process is peer review, not editorial review. Several reviews, particularly from strong-minded members, seemed unsympathetic to the authors' approaches, but were carefully argued, and it would have been improper to reject them. But where a review, especially a negative one, provides poorly explicated reasons, it seems to be fairer to seek another opinion. This we felt we could not do, but believe it was detrimental to the fairness of the process.

Conclusions

The confidentiality of the refereeing process means that this paper cannot be subject to some of the normal checks of academic research. And since some aspects of the refereeing process are idiosyncratic to individual editors, it is not possible to generalise our conclusions.

Nevertheless, the results illustrate several features of refereeing which are rarely formally analysed, and provide a research base for establishing principles of wide generality for sound refereeing.

We have not quite answered the claim of some rejected authors that some referees had not understood their papers. Because almost all of the rejection decisions were not unanimous, it is not surprising that authors felt that if one referee held a positive opinion, there were presumably others who would have agreed. We have certainly had the same experiences: trying to estimate one person's understanding of another's ideas must be a task for another paper. What we have done here is to show some of the reasons why there might have been misunderstandings, and to suggest ways in which the likelihood of misunderstanding might be reduced.

We had been told that the refereeing process can never be made foolproof. Here we have shown how the efficacy of the refereeing process rests strongly on the skills of the editors in matching papers with competent referees while working under significant time pressures, and also in making good value judgements about the quality of the reviews which are submitted. We have shown how errors can arise, but also how many potential problems can be reduced by more carefully constructed data bases or by allowing editors a little more autonomy to over-rule, or seek further opinions on, reviews they consider unsatisfactory. Our argument also suggests that there is a case for using permanent conference editors, rather than making use of people from the host city.

Nevertheless, there is little evidence to suggest that the 1999 procedures led to gross unfairness. Some allocations were only moderately good, and some reviews of poor quality, but these weaknesses were less apparent in cases where a paper was rejected than in general. This seems to have been fortuitous, thus emphasising the importance of careful prior organisation. No process is foolproof, but we have shown how the present procedures may be easily tightened.[1]

References

Clarkson, P.C. (1996). Report on the publishing process for MERGA 19. In *Guidelines for organising a MERGA conference.* Melbourne: Department of Science and Mathematics Education, University of Melbourne.

Sullivan, P., Owens, K, & Atweh, B. (Eds.) (1996). *Research in mathematics education in Australasia 1992–1995* No place of publication: MERGA.

Truran, J.M., & Truran K.M. (Eds.) (1999). *Making the Difference* (Proceedings of the 22nd annual conference of The Mathematics Education Research Group of Australasia Incorporated, held at Adelaide, South Australia, 4–7 July, 1999). Sydney: MERGA.

Truran, K.M., & Truran J.M. (2000). *Is the MERGA refereeing process improving the quality of Australasian mathematics education research?* Paper to be presented at the 23rd annual conference of The Mathematics Education Research Group of Australasia Incorporated, held at Fremantle, Western Australia, 5–9 July, 2000.

White, P. (1998). *Mathematics Education Research Group of Australasia Incorporated Membership Directory 1998* No publication details.

[1] *Acknowledgements.* We thank the MERGA Executive and the Conference Convenor, Brian Sherman, for inviting us to edit the Proceedings in the face of some strong opposition. We thank authors and referees for making our quite demanding job remarkably pleasurable. Finally, we thank Peter Brinkworth and Clive Kanes for so often acting as sounding boards in times of editorial crisis.