The reliability of peer reviews of papers on information systems

26 July 2003

Draft of paper published in the *Journal of Information Science* 2004 30: 2-11.

Michael Wood and Martyn Roberts
Department of Accounting, Law and Management Science
Portsmouth Business School
Locksway Road
Milton, Southsea, Hants, PO4 8JF, UK.
Tel: 023-92844168
Fax: 023-92844037
email: michael.wood@port.ac.uk

Barbara Howell
School of Information Management
Leeds Metropolitan University
Leeds LS6 3QS, UK.

Abstract

This paper analyses the reliability of the double-blind peer review systems used for submissions to the 2001 and 2002 UK Academy for Information Systems (UKAIS) conferences. The level of reliability found in the first conference was marginally lower than would be expected from a model based on chance. In the second conference the reliability level was significantly better, but still low. The paper explores some of the implications of this for the reviewing system, and suggests a model for assessing the impact of low levels of reliability.

Keywords

peer review, refereeing, reliability, journal review system, science management, knowledge management

1. Introduction

The main purpose of academic conferences and journals is to share ideas so as to assist in the growth of the academic discipline. To this end, a reviewing system is used to try ensure that papers presented or published are of adequate quality: papers are sent to reviewers - typically two - and the acceptance of each paper depends on the reports of these reviewers.

Such a reviewing system might be expected to enhance the quality of the papers accepted in three distinct ways:

1Most obviously, poor papers should be rejected, so the overall quality of the accepted papers should be enhanced.

2Comments are typically fed back to authors who are then asked to incorporate them into an improved draft.

3The fact that papers are reviewed might be expected to give authors an incentive to ensure that their work is of a high standard before submission.

Similarly, publishers may use two or three reviewers to help them decide whether to publish a book, and grant awarding bodies may use peer reviews to differentiate between successful and unsuccessful applications for research grants.

The peer review system is thus a very important part of the quality control system for academic knowledge. Its efficiency and effectiveness should be a matter of concern for all of us, both as academic workers whose careers are dependent on the system, and from the wider perspective of the development of academic knowledge.

A number of studies have been carried out to gauge the reliability and validity of the peer review process for academic journals in a variety of areas. This literature is reviewed in [1, 2, 3, 4].

The most widely cited study is probably Peters and Ceci [5], who resubmitted 12 articles to the psychology journals which had published them 18 to 32 months previously, after changing the names of the authors and institutions and a few other minor details. Only three of the articles were recognised, and eight of the remaining nine were rejected by the *same* journals that had

originally published them. Predictably, this article provoked a large number of diverse reactions, but it does seem to demonstrate the possibility that the reviewing process may, on occasions, be seriously flawed.

Cox et al [1] described the general level of agreement between reviewers of the same paper, as reported by the studies they discuss, as "low" in both the behavioural and the medical sciences. In a similar vein Daniel [2], in a study of the German chemistry journal, *Angewandte Chemie*, found that the "the observed extent of referee agreement must be regarded as rather unsatisfying" (p. 71): the values of the kappa coefficient of agreement (see Section 2 below) ranged from 0.12 to 0.25. More recently, a study by Rothwell and Martyn [3] on neuroscience papers found that for one journal the agreement between reviewers was no better than would be obtained by chance (kappa=0.1), and for another journal agreement was only slightly better than chance (kappa=0.3). Similarly, Weller [4] concluded, from a very extensive review of the literature, that "there is not a lot of agreement among reviewers" (p. 192).

All this evidence suggests that, in the contexts studied at least, the peer review system cannot be performing the first of the three functions listed above - that of preventing the publication of poor papers and facilitating the publication of good ones - because there is little agreement on which papers are poor and which are good. The consequences of this are potentially serious: Horrobin [6] suggest that the ineffectiveness of the peer review process may be responsible for a lack of progress in his field, psychopharmacology: in particular "the fact that in both diseases [depression and schizophrenia] the efficacy of modern drugs is no better than those compounds developed in 1950".

There is, then, considerable evidence that peer review is often unreliable, although, not surprisingly, the pattern varies from subject to subject, and journal to journal. Similar conclusions apply to grant applications: Hodgson [7] found that the agreement between the assessments of the same proposals by two agencies was "only fair" (kappa = 0.3) by one method of analysis, and "moderate" (kappa = 0.4) by another.

Agreement between reviewers is a measure of the reliability of the review process. There

is also a question about its validity. Suppose there is a "true" grade for the academic quality of each paper - either good or bad, but of course we do not know which. If two reviewers agree, can we assume that they agree because they have both managed to assess this grade correctly? Unfortunately we cannot assume this because there are several other possibilities. They may be assessing some irrelevant aspect of the paper (eg the quality of the printing or the diagrams), or they may both have made similar errors in assessing the quality of the paper (eg both failed to spot a statistical error), or the agreement may just be a matter of chance. Reliability does not necessarily indicate validity.

We have assumed so far that each reviewer produces an overall recommendation, and that all reviewers have the same status so that reviewers of the same paper should, hopefully, agree. The analysis in Sections 2-6 of the present paper is also based on this assumption. However, sometimes this assumption may not be justified: reviewers may be deliberately chosen to represent different viewpoints [4] (p. 199) leaving the editor with the task of synthesizing these differing perspectives. In this situation, there is no reason to expect reviewers to agree, so disagreement does not indicate a flaw in the reviewing process, and reliability cannot be assessed by agreement between reviewers. The question then arises of the reliability of the reviews from each perspective, and of obtaining two reviews from each perspective to assess this - which is probably not a practical proposition.

It is difficult to see what other approach besides peer review (in some form) could be used to assess validity. There is an extensive literature on assessing journals and the quality of papers in them [eg 8, 9]. This is a slightly different problem: it is possible to use retrospective measures of impact or quality for journals. One possibility here is citation analysis, which is obviously not feasible for assessing the quality of individual articles before publication. It is, however, possible after publication: Daniel [2] reports on a comparison between citation rates for articles published in *Angewandte Chemie* and for articles that had been rejected by this journal but published elsewhere - the citation rates for the accepted articles were roughly twice those for the rejected articles. The problem with this is, as Daniel acknowledges, that articles published in second

choice journals are unlikely to be as widely read, so this may be more of a self-fulfilling prophecy than a valid measure of article quality. Despite its faults, there seems little alternative to peer review.

The peer review process depends on the quality of the reviews, so instruments for measuring this are of obvious value: van Rooyen et al [10] reported favourably on one such instrument. This presumably leads to the possibility of evaluating reviewers and perhaps training them. This is clearly an important avenue to explore, although it is important to remember that even a reliable reviewing process may not deliver the results we might hope for in the long term. Would a peer review system have stifled the Copernican hypothesis?

The first of the functions of the reviewing system listed above (rejecting poor papers) requires reasonable levels of validity and reliability, but this is less obviously true of the other two. For the second function - feeding back comments - one critical issue is the extent of the agreement between reviewers' comments [11, 12]. As any recipient of two reviews of the same paper is likely to be well aware, there are often high levels of disagreement between reviewers here too: "in the typical case, two reviews of the same paper [submitted to American Psychological journals] had no critical point in common" [12]. The problem was not that reviewers disagreed, but that they picked up different criticisms. This leads on to a number of suggestions: training reviewers to be more thorough, deliberately choosing reviewers likely to disagree so that the comments authors receive are more varied, and increasing the number of reviewers so that more points are likely to be picked up. According to one of the commentators on the article by Peters and Ceci [13], the journal *Current Anthropology* solicited up to 15 reviews of each article for this purpose, and because "dependence on two or three referees is ... downright dangerous". If reviewers' comments are regarded as providing a sample of the population of possible points to be made about the paper, the problem is that of deciding how large the sample of reviewers needs to be to achieve a "reasonable coverage" of this population - the framework and methods developed by Wood and Christy [14, 15] could be used for this purpose.

The results of such an analysis are, however, likely to indicate a greater number of

reviewers than is possible under the present system. Fiske and Fogg [12] suggest authors should take responsibility for this process by getting their papers reviewed before submission by both friends and, more importantly, "enemies" or devil's advocates.

Taking a wider perspective, there are different conceptions of knowledge, which may have implications for the peer review system. If one rejects the idea of "truth" as a goal for knowledge, then the notion of validity of a review may take on a different complexion, and reliability may be less important. And with web journals, new systems for reviewing papers are likely to evolve [4]. One possibility would be to publish all submissions that pass an initial screening, and then add reader ratings or reviews (eg the systems used by the web bookseller, *Amazon* at www.amazon.co.uk, and the *Global Ideas Bank* at www.globalideasbank.org), and possibly citations, as time passes. Eventually, these might provide a better evaluation of each paper than the current peer review process, although Bingham et al [16] concluded that in their field (medicine) this is "no substitute for commissioned prepublication review".

The aim of the present paper is more restricted than this - to consider the reliability of the peer review process for recommending acceptance or rejection in a sub-discipline of management: the study of information systems. We present a statistical analysis of the reviews of the papers submitted to two annual conferences (2001 and 2002) run by the UK Academy for Information Systems (UKAIS). The review process was double-blind: reviewers were not aware of authors' identities (and vice versa). We go on to discuss some of the implications for authors, editors and the development of academic knowledge.

2. The reliability of the reviews of papers submitted to UKAIS, 2001

Reviewers at both conferences were asked to grade papers on a four point scale - accept, accept with minor revisions, accept with major revisions, and reject. This section analyses the data from the 2001 conference: we turn to the 2002 conference in Section 4.

There were 58 papers submitted that were reviewed by two reviewers. If the reviewing system were reliable and accurate, we would expect reviewers to tend to agree. Good papers

would get good reviews, bad ones would get bad reviews, but whatever the quality of the paper, the two reviewers should agree. Table 1 shows the actual picture.

Table 1: Number of papers submitted to UKAIS 2001 with each combination of grades from two reviewers (n=58)

	Accept	Accept/ minor revisions	Accept/ major revisions	Reject
Accept	*1*			
Accept with minor revisions	8	*11*		
Accept with major revisions	3	15	*3*	
Reject	4	7	5	*1*

If the two reviewers always agreed, all the papers would lie on the diagonal (figures in italics). Only 16 of the 58 papers (28%) lie on this diagonal - not a very impressive rate of agreement. In fact there was only one paper for which both reviewers recommended acceptance without amendments, and only one for which both recommended rejection. Even more strangely, there were four papers for which one reviewer recommended straight acceptance and the other outright rejection. This does not indicate a high level of agreement between these particular reviewers!

The 28% agreement rate ignores the fact that disagreement by one point on a four point scale is a relatively minor problem, whereas the four papers which received the rating "accept" from one reviewer and "reject" from the other demonstrated the sharpest possible disagreement. To clarify the notion of agreement between the reviewers, and to avoid setting an unrealistic standard (agreement on a four point scale), we collapsed the scale into two points - the first two

grades counting as good reviews, and the second two as bad reviews. (We tried a variety of other ways of analysing the data: all led to very similar conclusions.) With these definitions, 20 papers received two good reviews (the three cells in the top left of Table 1), 9 papers received two bad reviews and 29 received one good and one bad review. The reviewers agreed for 29 (20+9) of the 58 papers (50%), and disagreed for the other 29 (50%).

As noted in the introduction, the implications of agreement are unclear: the agreed verdict may or may not be valid. The implications of disagreement, on the other hand, are clear: if one reviewer one says the paper is good, and the other says it is bad, this is unhelpful because we can draw no clear conclusion (unless the reviewers are evaluating the paper from different perspectives as discussed in Section 1 - we are assuming this does not apply here). We cannot even conclude that the paper is of a type that draws mixed reviews: if another two reviewers had been chosen, then the paper might have received two good reviews. Disagreement between reviewers is definitely evidence for an unreliable, and so invalid, reviewing process, but agreement is not clear evidence for validity or reliability.

For this reason, we decided to focus on disagreement between the two reviewers of a paper, because the implications of disagreement are clearer than those of agreement between the two reviewers.

The disagreement rate in this case was 50%. This might not sound too bad. It is, however, necessary to put this 50% disagreement rate in the context of the worst performance that can reasonably be expected: ie the scenario in which the reviewers fail to read the papers and give their assessments at random. Imagine that, instead of reviewing the papers, reviewers were asked to toss a coin and then give a good review if it landed with its head uppermost, and a bad review if it landed tails uppermost. Probability theory then indicates that a quarter would receive a head (good review) from both reviewers, a quarter would get two bad reviews and a half would get mixed reviews. The proportion of papers on which the two reviewers disagreed would be 50% - exactly the proportion observed!

A slightly more subtle random model would take account of the fact that good reviews

were produced more frequently than bad ones: 69 of the 116 reviews were good. If reviewers were given a coin specially weighted to produce heads with a probability of 69/116 (59%), the expected results are given in Table 2, which includes the actual results for comparison.

Table 2: Actual results and results expected from a random model (UKAIS 2001)

	Actual proportion of papers (n=58)	Proportion expected from random model (p=69/116)
Two good reviews	34%	35%
Two bad reviews	16%	16%
Total agreement	50%	52%
Disagreement: one good & one bad review	50%	48%

Table 2 shows that the actual results are remarkably similar to this random model. In fact, there was a marginally *higher* level of disagreement in the actual results than is predicted by this random model (50% vs 48%), although the difference is far too small to be statistically significant. This suggests that the reviewing process was as unreliable as it could be.

Cohen's kappa [2] provides a measure of agreement between judges which is scaled so that the chance level of agreement corresponds to a value of zero, and complete agreement corresponds to a value of kappa of one. In this case

$$\kappa=\frac{50\%-52\%}{100\%-52\%}=-0.04$$

(The formula is simply the amount by which the actual agreement is better than chance, divided

by the maximum possible value of this quantity.)

The fact that the data is consistent with the assumption that the reviewers were making decisions by tossing coins does not, of course, prove that the reviewers were in fact doing this. (Another possible explanation is discussed in the next section.) *It does, however, suggest, that the process is no more consistent, and so no more useful, than this.*

One possible flaw in this conclusion is sampling error. We may have been unlucky in finding such a high level of disagreement between reviewers; with another sample of papers and reviewers the proportion of papers with disagreement between reviewers may be substantially less than 50%. The 95% confidence interval for the 50% level of disagreement between reviewers extends from 37% to 63% (using the standard formula for the confidence interval for a proportion). This suggests that the level of disagreement with this reviewing process is *at least* 37%.

3. Is the high level of disagreement due to differing standards set by reviewers?

An alternative to the assumption that the review process is a random one, is the assumption that some reviewers set higher standards than others, and that these reviewer standards are the main determinant of the reviews a paper receives. This would mean that there was little tendency for one paper to get similar reviews from two reviewers because the reviews depend on the reviewer not the paper. If this were true, we would expect some reviewers - those with high standards - to give bad reviews consistently, and others - those with low standards - to give good reviews consistently. In the extreme, there would be two types of reviewer - mean reviewers who always give bad reviews, and generous reviewers who always give good reviews. This hypothesis is different from the coin tossing hypothesis in that more laudable motives and procedures are attributed to reviewers; the effectiveness of the procedure, is, however, on a par with coin tossing.

Some of the reviewers reviewed a single paper, some reviewed two, some three, some four and one reviewed five. Clearly we cannot check this hypothesis with reviewers who

reviewed just one paper, and as the majority of the reviewers reviewed two papers, we have restricted this part of our analysis to this group. There were 37 reviewers who reviewed two papers, and of these

Proportion of reviewers giving two good reviews= 43% (16)

Proportion of reviewers giving two bad reviews = 19% (7)

Proportion giving one good and one bad review = 38% (14)

As before, it is helpful to compare this with a random model. Of the reviews given by these reviewers, 62% (46/74) were good and 38% bad. This leads to a prediction, based on a random model, of 53% ($0.62^2+0.38^2$) of these reviewers giving two similar reviews (two good or two bad). The actual figure, 62% (with a 95% confidence interval extending from 46% to 78%.), is higher than this, but not by a large margin (and 53% is within the 95% confidence interval so the difference is inconclusive from this point of view). There is some evidence that reviewers tend to have consistent standards, and this might be part of the explanation for the lack of agreement between reviewers of each papers - but this effect is small so it is unlikely to be the whole explanation.

4. The reliability of the reviews of papers submitted to UKAIS, 2002

We analysed the data on disagreement between the two reviewers of each paper for the next conference in the series, UKAIS 2002, in the same way. The results are in tables 3 and 4.

Table 3: Number of papers submitted to UKAIS 2002 with each combination of grades from two reviewers (n=68)

	Accept	Accept/ minor revisions	Accept/ major revisions	Reject
Accept	*4*			
Accept with minor revisions	15	*14*		
Accept with major revisions	3	12	*2*	
Reject	0	7	7	*4*

Table 4: Actual results and results expected from a random model (UKAIS 2002)

	Actual proportion of papers (n=68)	Proportion expected from random model (p=88/136)
Two good reviews	49%	42%
Two bad reviews	19%	12%
Total agreement	68%	54%
Disagreement: one good & one bad review	32%	46%

These results show a lower disagreement rate: 22 out of 68 papers, or 32% as opposed to 50% the previous year. The value of *kappa* is *0.30* - which is represents a "poor" level of agreement (Rothwell and Martyn, 2000: 1965).

A null hypothesis of random reviews predicts a mean disagreement rate of 46%, and a probability of 1.5% of getting 32% (22 out of 68) or fewer mixed reviews (using the Excel binomial function =binomdist(22,68,0.46,true)). This (one-tail) *p* value suggests that the results are not consistent with the random review hypothesis, whereas the previous year's results clearly were. On the other hand, 32% is still a high disagreement rate, indicating a moderately unreliable system.

There were two main differences between the reviewing systems in 2001 and 2002. First, fewer reviewers were used in 2002, which might be expected to lead to less disagreement. (The fact that relatively few of the reviewers assessed just two of the papers means that it is difficult to check this factor with the simple method of analysis in Section 3.) Secondly, the 2002 conference was organised by a different team, with different contacts, and doubtless different instincts about suitable reviewers for the papers.

It would be possible to analyse these differences further, but this would be of dubious value because another conference (2003 perhaps) might raise yet another set of factors to consider. Without a sample of conferences, or journals, sufficiently large and representative to use as the basis for statistical conclusions which would apply across academic sub-cultures, all we can do is note the following:

1The disagreement rates at the two conferences were significantly different; and

2Both disagreement rates - 50% and 32% - were large enough to raise serious doubts about the usefulness of the reviewing process.

The first point is in line with the "impression" of Fiske and Fogg (1990: 597) that "the basic unit was each editor and that editor's choice of reviewers", rather than the subject area. Each *academic sub-culture* is likely to have its own pattern of peer review reliability. And the second point indicates that in the two subcultures we looked at (and others reported in the literature), the peer review process *cannot* be relied on to yield valid assessments of papers.

This suggests that editors should analyse their subculture and take account of its properties when making decisions on papers. In Section 6 we outline a very crude statistical

model for this purpose. Before that we review some of the possible reasons for disagreement between reviewers.

5. Possible reasons for high levels of disagreement between reviewers

We have mentioned above (Section 3) one possible reason for high levels of disagreement - the possibility that some reviewers will expect higher standards than others. There are many other possibilities - for example:

1If the standard of the papers is very similar, so that there really is little to choose between them, we might expect less agreement between reviewers than if some papers were very obviously good, and others were very obviously poor.

2If papers and reviewers come from a variety of different academic perspectives, then agreement between reviewers is obviously less likely than if papers and reviewers are more homogeneous. (If the reviewers are *deliberately* selected to represent differing viewpoints they are, in effect, assessing something different and reviewer agreement ceases to be a useful measure of reliability - see Section 1.)

3If reviewers are not expert in their fields, their assessments may be more unreliable than their more expert colleagues.

4If different reviewers are using different criteria (eg some may be looking for originality, others may be focusing on links to existing knowledge, and others may be focusing on presentation) then they are likely to come to different conclusions.

Factors such as these mean that we would expect different levels of disagreement between reviewers at different conferences and for different journals.

It may be tempting to try to impose agreement by giving reviewers clearer guidelines (to deal with 4), or by restricting reviewers to those of a specified competence (3), or to those considered likely to agree with each other (2). However, it is worth reiterating that agreement between reviewers does not mean that the agreed assessment is valid, and there is an argument that disagreement is healthier for the growth of the discipline than agreement imposed by

inappropriate criteria. This is certainly likely to be true of agreement over the comments given to authors: there is little point in having two detailed reviews which say the same things.

On the other hand, imposing agreement is obviously a good idea if the first reason applies: if all the papers are of similar quality, they should obviously receive the same treatment.

6. A model for drawing inferences from peer reviews in a particular academic sub-culture

One of the papers submitted to the 2001 conferences received two bad reviews. It was rejected, although, in retrospect, this was perhaps unfair given the low reliability of the review process. Two negative reviews, however, tend to make a deep impression: it may be hard for editors to remember how unreliable the process is, and be prepared to over-ride the opinions of two experts. On the other hand, if the disagreement rate had been low - say 5% (instead of 50%) - then rejection would be far more reasonable. How should editors take account of reviewer disagreement rates in their subculture?

The first thing they should do is to calculate reviewer disagreement rates (d) and the overall proportion of good reviews (g). For the 2002 conference, for example, d was 32%, and g was 65%.

Consider an individual paper for which we have two reviews. Now *imagine* it has been reviewed by a large number of reviewers. The proportion of favourable reviews might be 100%, or 90%, or 80%, or 70% and so on. Different papers would draw differing proportions of favourable reviews. It would be nice to know where our paper falls on this scale. Unfortunately this is obviously not possible from just two reviews.

However if we simplify the situation by imagining that there are just three types of paper - *good papers* that receive 100% good reviews, *bad papers* that receive 100% bad reviews, and *mixed-verdict papers* that receive some of each - it is possible to build a model to give a rough idea of the proportions of papers in each category, which will give us a baseline against which to assess an individual paper. Call the proportion of good papers p_g, the proportion of bad papers p_b,

and the proportion of mixed-verdict papers p_m. Then obviously

$p_g + p_m + p_b = 1$.

All disagreements between two reviewers must stem from papers in the mixed-verdict category. Let's assume that the proportion of good reviews received by *all* the papers in this category is g. (This is a natural intermediate value between 0 and 1, and ensures the model has a feasible solution). If we choose two reviewers at random to review a paper in this category, the probability of their disagreeing is $2g(1-g)$, and the disagreement rate will be this probability multiplied by p_m:

$d = 2g(1-g)p_m$

Since there is a 100% probability that a review of a paper in the good category will be good, a probability g for the mixed-verdict category, and zero probability for the bad category,

$g = p_g + gp_m$

Solving these equations gives

$$p_g = g - \frac{d}{2(1-g)}$$

$$p_m = \frac{d}{2g(1-g)}$$

and

$$p_b = 1 - p_b - p_m$$

To take the second conferences as an example

$d = 0.32$ and $g = 0.65$

so

$p_g = 0.19$

$p_m = 0.70$

$p_b = 0.11$

The data from the conference is consistent with the assumption that 19% of the papers were "good", 70% were "mixed verdict" and 11% were "bad". Comparing these figures with Table 4,

we can see that the proportion of good papers is far lower than the proportion which obtained two good reviews (49%), and a similar result applies for the bad papers. The reason for this is that many of the papers with two good reviews are actually in the mixed-verdict category.

We can now use Bayes' theorem to work out what can be concluded about a paper from two good reviews. (Such a paper might be a good paper, or it might be a mixed-verdict paper, but it cannot be a bad paper because we are assuming these receive no good reviews.)

$$P(\text{paper mixed given two good reviews}) = \frac{q^2 p_m}{p_g + q^2 p_m}$$

Similarly,

$$P(\text{paper mixed given two bad reviews}) = \frac{(1-q)^2 p_m}{p_b + (1-q)^2 p_m}$$

Given the model, the inference from mixed reviews must be that the paper is in the mixed-verdict category.

For the second conference these probabilities come to

Prob(paper is good given two good reviews) = 0.39

Prob(paper is mixed given two good reviews) = 0.61

Prob(paper is mixed given two mixed reviews) = 1

Prob(paper is bad given two bad reviews) = 0.56

Prob(paper is mixed given two bad reviews) = 0.44

These results provides a warning against taking the reviews at face value. The probability of a paper which receives two good reviews actually being in the good category - meaning that any further reviews would be good - is only 39%. There is a 61% chance that it would actually turn out to be in the mixed-verdict category if more reviews were obtained. Similarly, there is a 44% chance that a paper with two bad reviews is actually in the mixed-verdict category.

These formulae can easily be adjusted to model the situation where we have three or more

reviews. For example

$$P(\textit{paper good given three good reviews})=\frac{p_g}{p_g+g^3p_m}$$

which is 0.50 for the 2002 conference. This is better than the two review result, but not by a large margin.

In practice a reasonable policy may be to publish papers in the good and the mixed-verdict categories, but not the bad category. This suggests rejecting papers with two bad reviews and accepting the rest. Under the assumptions of our model, the only error we can make here is rejecting a paper with two bad reviews which is actually in the mixed category. The probability of this error in the 2002 conference was 44%. We could reduce this error by using more reviews: the probability of this error is given by

$$P(\textit{paper mixed given n bad reviews})=\frac{(1-g)^np_m}{p_b+(1-g)^np_m}$$

If we wanted to reduce this error to, say 5%, in the 2002 conference, this equation indicates that we would need $n = 5$: ie we would need an extra three reviewers to check the papers with two bad reviews.

To take a contrasting example, if the disagreement rate at this conference had been lower - say 5% - the probability of a paper being good given two good reviews would have been 93%, indicating a much higher level of confidence in the verdict from two reviews. And if there were no disagreements ($d = 0$), the probability would have been 100%.

For the 2001 conference d=0.5 and g=0.59, which gives p_m = 1.03, which is clearly impossible. The problem is that the assumptions in the model cannot simulate a level of reliability worse than chance - as at this conference. If we use the chance level of disagreement (0.48), we get p_m = 1: ie all the papers must have been in the mixed category, and the probability of any

particular paper being in this category is 100% *regardless of what the reviewers say*. The reviewing process here was a waste of time from this perspective.

However, this is just a simple model based upon just three arbitrary levels of reviewer verdict of papers. Different assumptions would obviously lead to different conclusions, and there is no obvious way to validate any particular set of assumptions. The main purpose of this model is simply to demonstrate the possibility of high probabilities that two reviews may give a misleading verdict.

7. Conclusions

The unreliability of the peer review process in the two conferences we studied means that the process is little better than chance from the point of view of recommending acceptance or rejection of papers.

In a wider context, to the extent to which other reviewing processes show similar levels of unreliability, the progress of knowledge is likely to be hindered by the fact that the selection process is not effective at selecting the best papers. If doctors tested for diseases by tossing coins there would be an outcry, and yet a process with similar levels of effectiveness is routinely used in the academic world for vetting the claims made for the treatments doctors prescribe.

From the point of view of editors, it is important to note that reliability levels are often low, and are dependent on the particular academic sub-culture - our results showed significant differences between two conferences in the same series. The owners of each subculture should be aware of the reliability levels of their reviewers' verdicts, and should take account of this in their decisions. The model we presented in Section 6, allows us to convert a measure of reliability (the disagreement rate, *d*) to a rough estimate of the probability of making errors if we accept the verdict of two or more reviewers. If the probability of making these errors is considered unacceptably high, there are only two alternatives: either abandon the peer review process in its present form, or improve the reliability of the process - perhaps by using more reviewers for borderline submissions. (However, as we pointed out in Section 1, if reviewers are deliberately

chosen to represent differing viewpoints, disagreement between reviewers cannot be used to assess reliability so this model cannot be used. This does not, of course, mean that the reviewers are reliable, just that there is no available evidence.)

The main danger is that agreement between two reviewers may appear more conclusive than it actually is. For example, 19% of papers at one of the conferences received two bad reviews, but the model in Section 6 above suggests that, if further reviews were to be solicited for any of these papers, there is a 44% probability that at least some of these would be positive. A total of five reviews would be necessary to achieve the conventional 95% confidence level.

From the point of view of authors of articles, the implications are much simpler. Don't be depressed if your paper is rejected: it may be the luck of draw. Resubmit it to another journal - preferably one with good record in giving quick decisions.

From the point of view of the healthy growth and dissemination of knowledge, the picture is more confused because many rejected articles are subsequently published in other journals [4], but obviously after a delay and often in journals of a lower status [17]. We can envisage two extremes. The first is where there is only one acceptable journal in the field so rejected manuscripts cannot be published. The second is where there are so many acceptable journals that almost all manuscripts will eventually find a publisher. Low levels of reliability and validity in the reviewing process are likely to cause difficulties from the perspective of the credibility of the discipline at both of these extremes. At the first extreme, the arbitrariness of the selection procedure will lead to the rejection of good work and the publication of bad work. At the second extreme, this arbitrariness means that bad research is eventually likely to get through the filtering process. Most disciplines are likely to fall between the two extremes, but there is an argument that history is close to the first extreme and medicine is close to the second [4] (p. 196).

The peer review process does have two other important functions: feeding back comments to authors, and giving authors an incentive to improve their papers. Neither of these is necessarily undermined by the unreliability of the process, although the second of these may be thwarted if the illusion on which the peer review process depends is dented. The function of feeding back

comments to authors raises a very different set of issues, which we discussed briefly in the introduction.

Acknowledgement: We are grateful to Steve hand for his comments on an earlier version of this paper.

References

[1]Cox, D., Gleser, L., Periman, M., Reid, N., & Roeder, K. (1993). Report of the ad hoc committee on double-blind refereeing. Statistical Sciences, 6(3), 310-330.

[2]Daniel, H. D. (1993). Guardians of science: fairness and reliability of peer review (translated by W. E. Russey). Weinheim and New York: VCH.

[3]Rothwell, P. M. & Martyn, C. N. (2000). Reproducibility of peer review in clinical neuroscience - is agreement between reviewers any greater than would be expected by chance alone? Brian, 123, 1964- 1969.

[4]Weller, Ann C. (2001). Editorial peer review: its strengths and weaknesses. Medford, New Jersey: Information Today Inc.

[5]Peters, D. P., & Ceci, S. J. (1982). Peer review practices of psychological journals: the fate of published articles, submitted again. The behavioral and brain sciences, 5, 187-255.

[6]Horrobin, D. F. (2001, February). Something rotten at the core of science. Trends in Pharmacological Sciences, 22(2), 51-2.

[7]Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. J Clin Epidemiol, 50(11), 1189-1195.

[8]Jones, M. J. (1999). Critically evaluating an applications vs theory framework for research quality. Omega, International Journal of Management Science, 27, 397-401.

[9]Bonnevie, E. (2003). A multifaceted portrait of a library and information science journal: the case of the Journal of Information Science. Journal of Information Science, 29(1), 11- 23.

[10]van Rooyen, S., Black, N., & Godlee, F. (1999). Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. J Clin Epidemiol, 52(7), 625-629.

[11]Garfunkel, J. M., Ulshen, M. H., Hamrick, H. J., & Lawson, E. E. (1990, March 9). Problems identified by secondary review of accepted manuscripts. JAMA, 263(10), 1369-1371.

[12]Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. Am Psychol, 45, 591-598.

[13]Belshaw, C. (1982). Peer review and the Current Anthropology experience. Behav Brain Sci, 5, 200-201.

[14]Wood, M., & Christy, R. (1999). Sampling for possibilities. Quality & Quantity, 33, 185-202.

[15]Wood, M., & Christy, R. (2001). Prospecting research: knowing when to stop. Marketing Letters, 12(4), 299-313.

[16]Bingham, C. M., Higgins, G., Coleman, R., & Van Der Weyden, M. B. (1998, August 8). The Medical Journal of Australia internet peer review study. The Lancet, 352, 441-5.

[17]Cronin, B., & McKenzie, G. (1992). The trajectory of rejection. Journal of Documentation, 48(3), 310-317.