How good is this conference? Evaluating conference reviewing and selectivity

Harold Thimbleby
Department of Computing Science
University of Swansea, Wales
Harold@thimbleby.net

Paul Cairns
Department of Computer Science
University of York, England
paul.cairns@york.ac.uk

Peer reviewing of papers is the mainstay of modern academic publishing but it has well known problems. In this paper, we take a statistical modelling view to show a particular problem in the use of selectivity measures to indicate the quality of a conference. One key problem with the process of conference reviewing is the failure to make a useful feedback loop between the referees of the papers accepted at the conference and their importance, acceptance and relevance to the audience. In addition, we make some new criticisms of selectivity as a measure of quality.

This paper is literally a work in progress because the 2012 BCS HCI itself conference will be used to close the feedback loop by making the connection between the reviews provided on papers and your (audience) perceptions of the papers. At the conference, participants will generate the results of this work.

Peer review, selectivity, reviewer feedback, statistical modelling, audience participation.

1. INTRODUCTION

The "selectivity" of a journal or conference is the number of papers published as a fraction of the number of papers submitted; smaller numbers are supposed to indicate higher quality. A typical reasonable conference might have selectivity of around 25% to 30% whereas a good or "selective" conference might have a selectivity as low as or better than 10%.

Impact factor (IF), the average number of citations per paper published over the last two years, is also used as an estimate of quality; however, it has limitations, as, for example, reported by *Nature* [Editorial, 2005]. Impact factors are highly discipline-specific, citations per paper do not follow a normal distribution (so averaging citations is not the right way to measure impact). For conferences, the key decision as to whether to attend a conference is made before the proceedings are published let alone after the work has started to be cited. Thus, impact factor is not a useful measure of conference quality in the first instance.

Simkin and Roychowdhury [2005] make one of many clear cases against impact factor, while the UK House of Commons Science and Technology Committee have published a recent wide-ranging report, *Peer review in scientific publications* [Science and Technology Committee, 2011].

In contrast to IF, selectivity can be measured continuously and it is immediately available making it much more appealing for conferences. However, selectivity, too, can be misleading for various reasons. If a conference, e.g., ACM CHI, receives a huge number of papers but has limited resources to run the conference, the selectivity appears to be better (it decreases so supposedly quality thus goes up) even though the quantity and quality of *accepted* papers could be unchanged.

As a statistic, selectivity is often fed back to referees and authors, but it is unclear how either of these groups of people should constructively use this information in future. For example, should referees look to apply the selectivity statistic to their particular set of future reviewing jobs? And how might an author use the selectivity to increase their chances next time?

Perhaps the most telling counter to selectivity as a useful measure of quality is this: the UK national lottery has a selectivity of around 2% (if you aren't worried about how much you win) which is much better selectivity than even the very best

conferences, but this hardly makes winning any recognition of quality; then compare the lottery's selectivity with the UK's Engineering and Physical Sciences Research Council, which over 2010–2011 had a selectivity for proposals in the field of the Digital Economy ranging from 65% to 100%, depending on the type of call [EPSRC, 2011].

Finally, both selectivity and impact factor are frequently mistakenly used to imply particular papers published in high impact factor or highly selective outlets are good. If the measures mean anything, they refer to some aggregate quality of a conference (or journal), not to any particular papers. Indeed, papers with the highest citations are usually survey papers, whereas authors rather their original work to have the most citations. Our memory of good papers at a highly selective conference like ACM CHI may be more likely due to selective recall: a larger conference will have more papers we can select from to optimise our preferences, and a conference with multiple streams will put no pressure on us to attend papers we don't like.

So overall, selectivity has little meaningful content in terms of quality of the papers. This point is made more concrete in the next section where a plausible model of the statistics of reviewing distributions shows that natural (or even unnatural) variation between referees can lead to quite substantial deviations of perceived quality from "actual" quality. In fact, we will show that selectivity is not just misleading but frankly deceptive.

Yet it is important to find a robust measure of conference quality. While a reader of a journal can easily decide whether to read an issue of a journal or not, or even just read the particular paper they are after without reading the entire journal (which is of course very easy with electronic publishing), the attender of a conference is in a different position. Before attending a conference, the participant has to make a major commitment to attend, to find the funds to travel, register and for accommodation — and commit the time away, maybe a week plus jet lag recovery. Attending a conference is an investment, and a participant needs something more indicative of quality than a number depending on the quantity of papers not accepted! And of course there are national schemes such as the UK's Research Excellence Framework that try to assess quality for funding purposes.

Selectivity purports to measure conference quality but referees try to measure individual paper quality. Unfortunately, referees don't do anything as simple as just measuring the quality of a paper; for example, because they want a paper to be accepted (or rejected) they may give it a very high score (or a very low score, respectively) to try to ensure the final average score is closer to what they want. In some refereeing systems, referees are asked to specify their expertise; and this is generally inaccurate: weak referees tend to overrate their expertise, and expert referees tend to underestimate their expertise [Kruger & Dunning, 1999; Ehrlingera, *et al*, 2008] — the Dunning-Kruger effect. Worse, a definite referee (who really wants to accept or reject the paper) may believe and claim they have a higher expertise than they do, because their certainty eclipses their self-expertise assessment.

Often referees delegate the work of refereeing to students [Science and Technology Committee, 2011], who may be unaware of relevant parts of the literature. Often referees (in our experience) seem to prefer scoring whether they like a paper, rather than assessing its scientific validity. In HCI itself, a very diverse discipline, these and other problems are exacerbated because it is hard for any individual to know the specialist research background of the paper well enough [Thimbleby, 2004].

Like selectivity, referee scores have little meaningful content, yet they too are treated as both significant and important. An author's career progression may depend on acceptance, the dissemination of good work (and the non-dissemination of poor work) depends on reliable refereeing, "best paper prizes" need to be awarded fairly, and refereeing is also used to make funding decisions for research proposals.

What can be done to improve the situation? Actually, nobody knows. This is because there is very little systematically collected data on, first, the reviews produced by referees and, crucially, the relationship between the reviews and the paper quality.

What is known is that in any situation where people are to improve there needs to be useful feedback. In the conference reviewing situation, the goal must be to select high quality papers where quality can be both audience perception at the conference and long-term relevance of the work. In all conferences, there is no systematic attempt to link the reviews referees produce to the quality of the papers as perceived by the audience. Reviewers are not alone in being unable to connect performance to feedback. Kahneman [2011] raises exactly this issue in relation to investment bankers and psychotherapists. Without correct and meaningful feedback, reviewers are likely to perform little better than chance and rely on surrogate approaches to reviewing that may have little bearing on the situation in hand.

This paper identifies one approach to solving the feedback problem by proposing, with agreement of the conference programme committee, to carry out a data gathering exercise in the course of this HCI conference. The data will address the two gaps in knowledge being first understanding the systematic variations in reviews and also relating the reviewer data to the quality of the papers as perceived by the audience. Our data will strongly depend on you, the audience, but we hope to substantially inform the reviewing process at least for this conference. In our presentation of this work, we hope to present some preliminary findings.

2. ANALYSING REVIEW SCORES

In order to better understand the issues with selectivity, we describe here a simple model that reflects our current best understanding of the refereeing process. There are necessarily some simplifying assumptions but all are plausible and moreover whilst the particular parameters and details may vary, we have tested many version of this model and the results are broadly similar. This is to say, there is nothing particularly idiosyncratic about this model nor has there deliberately been an attempt to devise a model that favours our view. Indeed, our view has largely arisen from thinking about models of this sort.

The first assumption is that quality of a paper is assessed by a single numerical value. This is in accord with many conferences where referees are asked to provide their confidence in a review on a scale from –2 to +2 or a scale of 1 to 10, etc. It is a hugely reductionist approach to quality, which could be considered anathema to the whole meaning of quality [Pirsig, 1974] nonetheless it is a well-established operationalisation of quality in most conferences. For simplicity, quality for the purposes of this paper is represented by a value between 0 and 1, where 1 is top quality and 0 is no quality at all. The idealised quality score of a paper q is called here the *true quality*.

The first question, then, is how quality varies across papers. It has been shown that quality follows a Zipf distribution [Anderson, 2009]. That is, when papers are ranked by quality, the quality of the paper, q, is a power law of the rank, r, thus: $q = r^{-p}$ where p is some number usually between 0 and 1. The Zipf distribution is a long-tailed distribution where there are a few very exceptional quality papers but where there is a long tail of papers with slowly declining quality.

The third assumption is the selection criterion. Here we have used a simple threshold argument that when the referees' scores average above a certain threshold then the paper is accepted. This is a simplification over what is normally done where borderline cases are considered more carefully but anecdotal report suggests that even allowing for this careful consideration it does seem to broadly hold that when the threshold of acceptance is passed then the paper is accepted. The threshold is set for us at 0.5 indicating that, on average, the paper quality is judged by referees to be more good than bad.

Using these assumptions, we can now describe the model. The Zipf parameter has been set to p=0.15 so that over a conference receiving 1,000 submissions, about 100 (actually 101) papers pass the acceptance threshold. Thus the conference ought to have selectivity of 10%. Moreover, this gives a long tail of papers that fall around the 0.4 quality level, which is just below the level of acceptance. This also seemed a reasonable test of the model as there are often papers that, for one reason or another, are considered as decent papers but not quite at the threshold for acceptance. It is rare to see a large set of very poor papers submitted to a conference. Figure 1 shows the distribution of a paper's true quality against the paper's rank based on true quality.

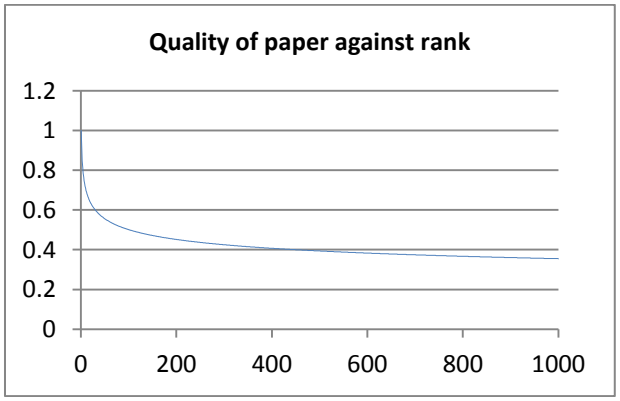

Figure 1: *The quality of a paper against its rank forms a Zipf distribution with parameter p. Here p=0.15.*

If referees were fully reliable in assessing papers then the average referee score would produce (or strongly correlate with) the true quality value. Hence only good papers would be accepted, and all good papers would be accepted — or, more usually, all the best papers down to some quality threshold set by the capacity of the conference to present them. However, in practice, referees see through a glass darkly and are only able to produce a score that is an imperfect reflection of the true quality. This is in part why we use several referees in order to effectively "sample" the space of quality scores and hence (hopefully) provide a more accurate, or at least a more defensible, estimate of the true quality. Even then there is some noise.

We have introduced noise into the referee scores by saying that referees vary from the true quality of

the paper according to some normal distribution. This is a plausible assumption given the theoretical basis for the normal distribution (Howell) but actually is not known for sure and we hope to address this with further work.

The noise is therefore modelled as normal distribution with a mean of 0 and a standard deviation of *s* and the noise is added to the true quality of the paper to produce a particular set of referees' averaged assessment of the quality of the paper. The parameter *s* needs to be decided on so in order to reflect modest variations, *s* was taken to be 0.1. This corresponds to 67% of refereed quality scores are within 0.1 of the true quality: so a 1 mark out on a 10 point scale or half a mark out on a five point scale. This seems pretty accurate refereeing in our experience.

The question then is, what is the selectivity of the conference based on this model? The noise is a random variable and so produces different values each time it is recalculated. A typical result of the model is given in table 1.

Table 1: *Contingency table of true quality decisions against refereed quality decisions.*

	Referee Accept	**Referee Reject**	**True totals**
True Accept	*72*	*29*	*101*
True Reject	*158*	*741*	*899*
Referee totals	*230*	*770*	

In general, the models produce a selectivity of between 20% and 25% with typical values being around 23%. This is far higher than the true selectivity of 10%. In addition, only a proportion of the true papers are accepted. In the example only 72 of the original 101 papers are accepted to the conference. Moreover, only 72 of the total of 230 accepted papers giving that only 31% of the accepted papers would actually meet the quality criterion.

Variations from this model produce consistent results, with selectivity being a lot higher (that is, double in value) than the true quality would suggest and moreover the problem of failing to accept good quality still appears. It could be argued that 72% is a good rate for publishing good quality work and some marginal quality work alongside. However, analysis also shows that in this specific example, 15 of the accepted papers are from the bottom 250 papers as ranked by true quality and 6 from the bottom 100. Thus, poor papers, not just marginal quality papers, are being mistakenly selected.

An easy attack on the model is the arbitrariness of the noise parameter, *s*.

Halving *s* to be 0.05, which corresponds to 95% of refereed scores being within 1 point of the true quality on a 10 point scale naturally gives better results. Now selectivity is typically 13% and 80 or so of the high quality papers are selected. However, this produces a very narrow variation in the overall referee scores, which does not reflect our experience of reviewing. Even so, in this case, only 62% of the accepted papers are in the high quality set.

Without proper data, what *s* should be is entirely moot. And that is another reason to collect data on this topic.

Overall then, a simple but plausible model suggests that selectivity is not an accurate reflection of what might be termed the true quality. Indeed, the true quality papers are only a modest proportion of all papers accepted and that, in some cases, even the worst papers are included for publication. This is clearly not the goal of a conference as normally conceived. **Selectivity is not only meaningless it is deceptive**. What is needed is a better understanding of how refereed scores relate to people's perceptions of papers. Only when that is known will it be possible to understand the problems of peer review, perhaps using models like this, and then also allow a feedback process that would help reviewers improve.

3. THE PROPOSAL

It is unconventional to propose in a paper the project that will collect the data that will form the basis for the presentation of the paper! It is for this reason that the BCS HCI 2012 conference organisers suggested we present our work in alt.hci (http://hci2012.bcs.org/calls.html#alt).

We now outline our current plans. In the talk (if referees dare to accept this paper), we will first present the data we will have gathered to that point and talk about the data gathering that will be still on-going at that point.

The starting point for our project is to gather the referees' scores for papers submitted to this conference. This will do two things. First, it will help to see the underlying distribution of quality of papers. Whilst this cannot give definitive answers to the distribution of the true quality, it can at least give indications of how the true quality might by distributed. Secondly, it will allow us to quantify the

natural variation seen between referees, both in relation to each other and in relation to the underlying true quality.

This data will be available in the advance of the conference and we will present it framed by the model above to structure the discussion.

The key step, though, is to close the loop and find out how the audience perceives the quality of the papers. This can then be related to the refereed scores.

One problem in a conference is that audiences are also swayed by their perceptions of the speaker and of the talk. Thus, a charismatic speaker may make even quite modest work seem very appealing whereas a more substantial and better quality piece of work may be presented in stultifying way that eclipses the underlying quality of the work.

Our idea is to provide conference audience members with a questionnaire to fill out after every talk that they attend. This questionnaire will encourage people to partition their perceptions of the talk from their perceptions of the paper. It may also be possible, if the returns are sufficient, to do analysis based on covariance to account for any halo effects given by a good speaker.

The data will be being gathered at the conference so we cannot promise to present anything on this though the intention is to proceed with data analysis whilst there. (Hopefully our presentation will be scheduled towards the end of the conference!)

A further measure of quality is citations of the work in future, though even citations (bibliometric data) are controversial measures of quality (consider that some seriously faulty work may have very high citation counts, as in cold fusion). Obviously this cannot be done at the 2012 BCS conference itself nor indeed in time for next year's conference but this is something we plan to revisit perhaps year on year and report back any useful findings at a later date.

(The closest prior work we know with similar concerns to ours is [Perneger, 2004], which compared citations to web hits, but this failed to relate either to referee scores.)

4. DISCUSSION & CONCLUSIONS

This paper has made three key points: (i) selectivity is misleading (ii) refereeing is uncalibrated, and (iii) we need to collect data. These points are inter-related, and we envisage that improvements to global quality measures (which is what selectivity purports to be) and to individual quality measures (which is what referees purport to provide) will become apparent as data from presentations is analysed. Hopefully, the analysis will lead to insights that can be used to improve the refereeing and selection process.

This paper has specifically addressed the problem of selectivity statistics for indicating the quality of a conference. The modelling shows that even quite simple models based on plausible assumptions show a distinct lack of correspondence between selectivity and the quality of papers. However, the model's assumptions may be plausible but they are unproven and untested and so we propose with this paper at this conference to establish some better empirical underpinnings for models of this sort.

Note though that this proposal suffers from a selection bias: we cannot evaluate papers that were rejected and hence are not going to be presented to any audience. Ironically this is one of the problems of selectivity: the fraction of rejected papers says almost nothing about the conference quality for the audience or for a reader of any accepted paper.

Selectivity is only part of the problem, though. The main problem is that peer reviewing is not and cannot be adequately taught while there continues to be no opportunity for formative feedback. Our proposed data gathering exercise offers the first step towards providing the necessary feedback by seeing what (if any) is the relationship between referee views (as represented in their scores) and audience perceptions (as represented in the evaluations). Both of these are imperfect but given the much larger sample represented by the audience of a paper compared to the number of referees, it is likely to be a much better indicator of the overall quality of a paper than the view of around three referees. On the other hand, referees are (generally) selected for their expertise (and spare time), whereas the audience is self-selected for a variety of reasons, including the desire to learn — almost the opposite of referees!

There is the caveat of course that we may all be so ossified in our reviewing practices and prejudices that we may fail to improve as a consequence of feedback. To some extent, we expect this from the infamous Dunning-Kruger Effect [Kruger and Dunning, 1999] whereby if a person does not have sufficient knowledge of a domain in which they are working, not only do they produce bad work but feedback is useless in helping them to produce better work because they do not understand the meaning of the feedback either.

It should of course be noted that true quality as described in this paper is something of a fiction but,

like good fiction, it is not without verisimilitude. Quality is essentially a multidimensional concept drawing on the methods used, the findings obtained, the writing style, and the relationship to the body (or bodies) of existing work. Thus, any reduction of quality to a one dimensional scale must necessarily abstract some of this complexity. However, this is the constraint under which most reviewing processes expect referees to work. It may be that quality should be acknowledged as multidimensional and that a paper need only pass the threshold on say 3 of 5 dimensions to be considered acceptable. But this begs the question of what those dimensions should be and whether referees can accurately assess them, and indeed whether the thresholds should be combined linearly. And there would still be the issue of how referees learn to produce better reviews.

In summary, this paper is clearly specifying the problems of reviewing and selectivity that apply to conferences generally (and to journals) and to this conference in particular. We therefore aim to at least start learning from this conference by gathering necessary data that until now has never been collected.

3. REFERENCES

Anderson, T. (2009) "Conference reviewing considered harmful," *ACM SIGOPS Operating Systems Review*, **43**(2):108–116.

Editorial (2005), "Not-so-deep impact," *Nature,* **435**:1003-1004, doi:10.1038/4351003b.

Ehrlingera, J., Johnson, K., Banner, M., Dunning, D., Kruger, J. (2008) "Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent," *Organizational Behavior and Human Decision Processes*, **105**(1):98–121.

EPSRC (2011) Research Proposal Funding Rates 2010-2011. www.epsrc.ac.uk/SiteCollectionDocuments/funding/FundingRates1011.pdf (14th June, 2012)

Kahneman, D. (2011) *Thinking Fast and Slow.* Allen Lane, London.

Kruger, J., Dunning, D. (1999) "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence leads to inflated self-assessments" *J. of Personality and Social Psychology.* **77**(6):1121–1134.

Perneger, T. V. (2004) "Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ," *British Medical Journal*, **329**(7465):546–547. doi: 10.1136/bmj.329.7465.546

Pirsig, R. M. (1974) *Zen and the Art of Motorcycle Maintenance.* William Morrow, New York.

Science and Technology Committee, House of Commons (2011), *Peer review in scientific publications*, Eighth Report of Session 2010–12, HC 856, The Stationery Office Limited, London.

Simkin, M. V., Roychowdhury, P. (2005) "Copied citations create renowned papers?" *Annals of Improbable. Research*, **11**(1):24–27.

Thimbleby, H. (2004) "Supporting Diverse HCI Research," *Proc. British Computer Society HCI Conference*, Leeds, UK, 6-10th September, 2004, 125–128. Research Press, London.