Expertise and the Wisdom of Crowds: Whose Judgments to Trust and When

Matthew B. Welsh (matthew.welsh@adelaide.edu.au)
University of Adelaide, North Tce
Adelaide, SA 5005, Australia

Abstract

The Wisdom of Crowds describes the fact that aggregating a group's estimate regarding unknown values is often a better strategy than selecting even an expert's opinion. The efficacy of this strategy, however, depends on biases being non-systematic and everyone being able to make a meaningful assessment. In situations where these conditions do not hold, expertise seems more likely to produce the best outcome. Amateurs and professional judgments are examined in a subjective domain – reviews of shows from an Arts festival – asking which group provides better information to the potential theatre-goer. In conclusion, while following the crowd produces good results, where a smaller number of reviews are available, taking expertise into account improves their usefulness and discrimination between shows.

Keywords: Expertise, Wisdom of Crowds, subjective judgment.

Introduction

When making decisions between diverse options, we often do not have sufficient time or resources to conduct the sorts of thorough analyses recommended by decision analysts (see, e.g., Newendorp & Schuyler, 2000). Instead, we rely on simple rules to greatly reduce the complexity of our decision making while maintaining as much quality as possible (Gigerenzer & Todd, 1999). Perhaps the simplest such rule is: if someone recommends option A, then I will select option A.

This approach, of course, requires that you have some idea of whether or not you should trust the opinion of the person offering it, which is easy when it is a person you know but more difficult when you are forced to rely on the opinions of strangers – as is often the case.

As an example, consider a person's decisions regarding what to spend his/her entertainment budget on. While they could wait and hope that their friends will go to see all of the various shows that they were interested in, more often, they will have to rely on reviews from either professional reviewers or sites such as "Rotten Tomatoes" that aggregate amateur review data. In either case, the criteria on which the reviewers have provided their rating is generally unknown to the people using the information.

The question, then, is how to make the best use of the available information – from both professional and amateur reviewers – in order to make informed decisions about the quality of entertainment on offer.

The Wisdom of Crowds

The wisdom of crowds describes a well-known effect first discussed by Galton (1907) and more recently repopularized by Surowiecki (2004). The observation is simply that, when making decisions under uncertainty, the median or mean estimate of a crowd is often a better predictor than the estimate of a randomly chosen individual – even an expert.

This initially surprising observation results simply from the underlying mathematics of the problem. If any biases or errors in people's estimates are independent, then they will tend to be in random directions and thus, when averaged, will be removed. This has allowed researchers to demonstrate that even having the same individual make an estimate twice and averaging those values can produce better estimates – so long as some degree of independence can be established between the two estimates (Herzog & Hertwig, 2009; Vul & Pashler, 2008).

For the wisdom of crowds to work, therefore, one needs to be considering a domain in which biases in people's judgments are not systematically related to those of other people. If this condition is met, then one expects that averaging the judgments of a group regarding the quality of a particular show would provide a better estimate of how much you will enjoy it than relying on the advice of any single reviewer.

Expertise

By comparison with the wisdom of crowds, expertise is a harder creature to pin down. While we all have an implicit understanding of what expertise is, actually defining it proves surprisingly difficult (see, e.g., Shanteau, 2002; Weiss, 2003) and people commonly confuse it with simple length of experience (Malhotra, Lee, & Khurana, 2005).

Despite this, given that we know there is such a thing as expertise and that people are employed on the basis of this to provide expert advice, it would seem reasonable for us to expect that this advice will be valuable – more valuable, at least, than a non-expert's judgment.

Decision Criteria

An important question, which should be asked before continuing, relates to the decision criteria being used. This is important as, when we ask a question, we can only receive meaningful responses if the person understands and answers the question we have asked. In the case of reviews of entertainment, then, what is the question that is being asked?

The difficulty here is that expert and non-expert reviewers may be answering different questions. Experts might be answering the question – how much artistic merit does the show have? Non-experts, by comparison, may be answering the simpler question – how much did you enjoy the show. In both cases, the judgment is subjective and dependent on the

reviewers personal tastes but, in the first, it is also being judged against taught norms of quality.

A secondary concern is the fact that most reviews are undertaken on an absolute scale, whereas people are far more comfortable and more accurate making relative judgments (see, e.g., Stewart, Brown, & Chater, 2005; Stroop, 1932). Given this, we need to be cautious in interpreting what a reviewer may mean by any given review.

This Study

In this study, reviews of entertainment will be analyzed in order to determine how a person could best use the available information to select a show to attend. It thus overlaps significant with problems such as the Netflix Prize (Bennett & Lanning, 2007) but is approached from a psychological rather than machine learning stance – that is, incorporating concepts such as expertise and considerations of *why* we have the data we do and how this should affect its use (for further discussions of this, second, point, see, Welsh & Navarro, 2011; Welsh, Navarro, & Begg, 2011).

Method

The data sets selected for analysis consisted of reviews of acts performing at the 2011 Adelaide Fringe Festival – a large, "unjuried" Arts Festival held annually in Adelaide, Australia. Being an unjuried festival, any act is free to register to perform without being selected by the festival's governing body. As such, the quality of performances is (presumably) more variable than would be observed in a juried festival where acts must convince the festival's jury of their quality before registering.

Given this, selecting a quality show to attend from the hundreds (750 in 2011) on offer becomes a difficult task in the absence of reliable indicators of quality. To this end, two databases of reviews were acquired: first, the Adelaide Fringe's summary of published, professional reviews from newspapers and news websites – labeled simply "Fringe" hereafter; and, second, the database from BankSA's "Talkfringe" website which allows anyone to register and post reviews of any Fringe shows that they have seen.

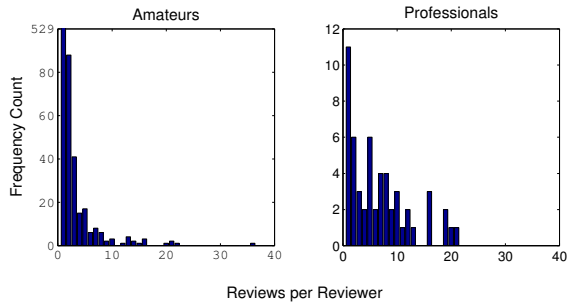
All of the Talkfringe reviews use the same 1 to 5 'Star' rating system (with half stars). The professional reviews, however, were in a variety of formats. To maintain comparability, therefore, only professional reviews that used a 5-star rating system were included in the analyses.

Data Characterization

The Fringe database records 365 reviews in the required 5-star format, made by 54 reviewers – an average of 6.8 reviews per reviewer. By contrast, the Talkfringe database contains 1436 reviews made by 731 reviewers. Figure 1 displays this information as a histogram of reviews per reviewer for the Amateurs (Talkfringe) and Professionals (Fringe) separately. Between the two databases, reviews were obtained for a total of 420 shows, with each being reviewed an average of 4.3 times.

Looking at Figure 1, one sees that both subplots seem to display similarly shaped distributions – a decay function of some type. The figure is, however, somewhat misleading as the y-axis of the Amateur subplot is displayed as if the highest count was 100 when, in fact, it was 529 (as indicated by the high value on the y-axis).

Figure 1. Histogram of number of reviews per reviewer by reviewer group. Note: Amateur y-axis is non-linear at top.



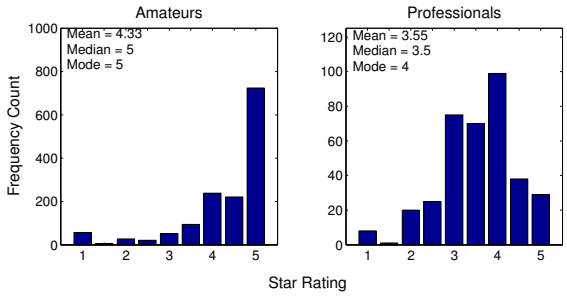
That is, while only a modest proportion (12/54) of the professionals reviewed only a single show, the majority of amateurs (529/735) did so.

Results

Indirect Comparisons

As an initial approach to the question of whose reviews should be trusted, the distributions of star-ratings within each database were compared. Figure 2 shows the histograms of this data.

Figure 2. Histogram of Star ratings by reviewer group.



Looking at Figure 2, one sees that the two distributions differ significantly from one another, as confirmed by an independent samples t-test, $t(1799) = 13.9$, $p < .001$, Cohen's $d = 0.81$. The Amateurs display something close to an exponential distribution of star-ratings, with a median and mode at 5 and a mean of 4.33, while the professionals display something closer to a Gaussian, with a mean and median around 3.5 and a mode at 4. This raises questions about the discriminability of Amateur reviews – that is, whether seeing a 5 star review from an amateur allows you to conclude anything meaningful about that show.

There are, however, alternate possible explanations for this pattern of responses. The first is that amateurs tend to be less discriminating in their tastes than the professional

and, thus, enjoy shows more. The second, however, is a selection effect – while professionals are told which shows to attend and write reviews of all of the shows that they attend, amateurs choose shows that they think they will like and are less likely to write a review unless motivated by particularly enjoying or disliking the show. Given that more popular shows attract greater audiences, and assuming a positive relationship between quality and popularity, this will tend to result in large numbers of high-star reviews for popular shows and relatively few reviews of any sort for less popular shows.

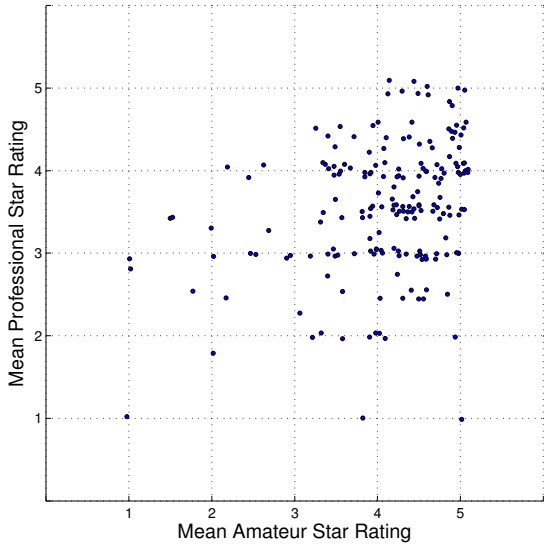
Based on this reasoning, one could assume that any show that has multiple, high-star reviews from amateur reviewers is likely to have been a popular show.

Direct Comparisons

The above discussion considers only the distributions of star ratings, rather than those instances where we have reviews of the same show made by both amateur and professional reviewers. An examination of the two databases revealed that, of the 420 shows, 191 of these were 'shared'; that is, had been reviewed by at least one member of each reviewer group.

Looking only at these 'shared' shows, the difference between the professional and amateur groups (3.59 versus 4.33) is almost exactly the same as for the full dataset (3.55 versus 4.33) and remains significant by a paired samples t-test, $t(1231) = 11.2$, $p < .001$, Cohen's $d = 0.79$.

Figure 3. Scatterplot of mean amateur versus mean professional review for all 191 'shared' shows. NB – some jitter has been added to the points to reduce overlap and facilitate display.



Despite the removal of over 200 shows that lacked a rating from each group, a consideration of only the overlapping shows still contains the majority of the review data as these 191 shows attracted 1233 of the total 1801 reviews and a comparison of the distribution of star ratings within this group with that for the complete datasets shown in Figure 2 revealed no noticeable differences. Figure 3 plots the mean reviews provided by each group for each show against that calculated from the other group.

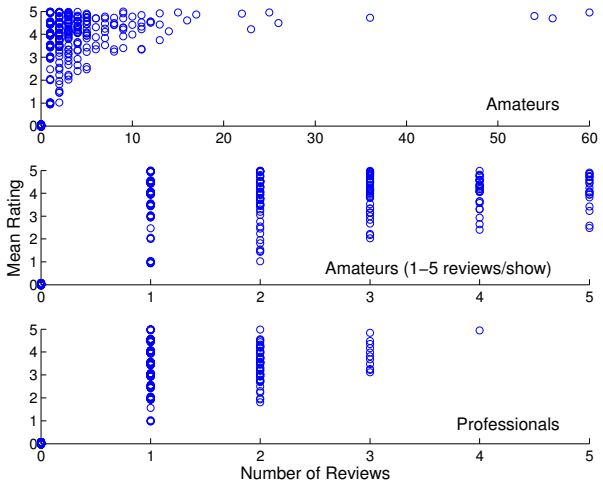
Looking at Figure 3, one can see that the relationship between the amateur and professional reviews is positive, but not particularly strong – confirmed by a correlation $r(190) = 0.32$, $p < .001$, indicating significant disagreement between the two groups on the quality of shows.

A closer examination of the figure reveals that a partial explanation for the poor correlation may be restricted range – with relatively few datapoints in the lower left quadrant. Again, this is likely to reflect selection biases, with all type of reviewers more likely to attend and review popular shows – which, in turn, are likely to be of higher quality.

Quality by Popularity

Given the data above, what can we say about how a person should go about selecting a show to see? As noted above, there is an assumption that higher quality shows are more likely to become more popular and that the number of reviews can be used as a proxy for popularity. This means that we can compare the star-ratings for shows of differing popularity to see how these variables interact. Figure 4, below, plots show star-ratings against number of reviews for all 420 shows contained in both databases.

Figure 4. Scatterplots of number of reviews (show popularity) versus mean rating (show quality) for Amateur and Professional reviewers. NB – some jitter has been added on the y-axis to facilitate display.



Looking at Figure 4, one sees that the mean ratings of shows that received low numbers of reviews vary quite significantly – indeed for shows with only one or two reviews, the mean ratings are fairly uniformly distributed across the 1-to-5 range.

For shows with higher numbers of reviews, however, one sees a striking pattern emerge – as the number of reviews increases, so does the *minimum* mean rating that that show

received. Comparing the bottom two subplots, one sees that this pattern emerges early in both the amateur and professional reviews; no show with 3 or more reviews averages less than a 2-star rating.

Looking across the top subplot of Figure 4, one can see this predictive power continues for higher numbers of reviews: no show with 6 or more reviews was rated lower than 3 star (on average); no show with 14 or more reviews was rated lower than 4 star (on average); and the 7 shows that were reviewed by 25 or more people all averaged at least 4.5 star reviews.

This would seem to confirm the prediction that popularity and quality are, in fact, linked and suggest that an appropriate strategy for selecting a quality show would be to select one that many people have reviewed – even without reading those reviews.

Expert vs Non-Expert Reviews

A final question to be addressed is that of expertise. While we have, above, divided reviewers according to whether they are Professional or Amateurs – and assume that this reflects some difference in expertise (in reviewing shows) – the data afford us some scope to test this assumption.

Looking once more at Figure 4, for example, one can see a suggestive pattern in the comparison between the Amateur and Professional results – where the speed at which the predictive multiple reviews increases seems greater for the Professional. That is, having had multiple Professional reviewers attend a show may be a better indicator of quality than having had the same number of Amateurs review it.

A more important question, however, is whether we can establish that expert reviews are *better* than non-expert reviews. The difficulty, of course, is in determining how we measure the quality of a review – after the fact and in the absence of any objective standard. A simple wisdom of crowds approach would suggest that we use the median or mean review from all reviewers as the standard but this runs into the problem of non-discriminability in the amateur data where too many shows will all be rated 5-star.

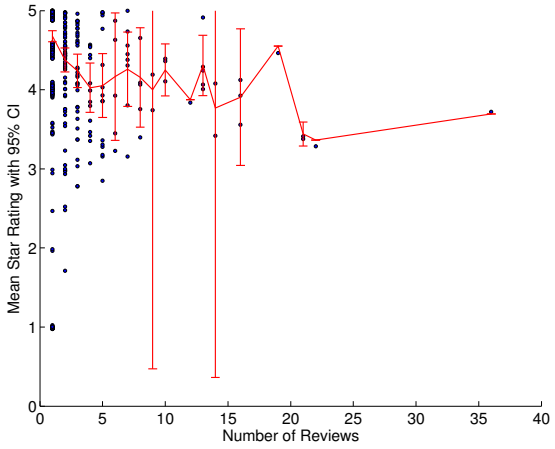
There are, however, at least two methods of using the current data to shed light on the relative usefulness of professional and amateur reviews in selecting a good show.

Measuring the Expertise of Amateur Reviewers

The first of these involves a comparison of the differences within the two groups. For example, it seems a reasonable assumption that those Amateurs who review more shows become more expert in doing so. The same relationship, of course, is less likely to hold in the Professional reviewers as the assumption is that these people have significant previous experience that is not available to us through the data set; and which is likely to outweigh any effect of the relatively few reviews they made during this event. Given the above, it seems necessary to restrict this discussion to differences within the Amateur group.

What then are the differences between the more and less 'expert' amateurs – that is, between those who posted many rather than few reviews. Figure 5 thus plots number of reviews per amateur reviewer against star ratings.

Figure 5. Scatterplot comparing number of reviews to mean star rating (amateurs only). 'Jitter' has been added to the data along the y-axis to prevent datapoints overlapping The red line shows the overall mean for each group of reviewers.



Looking at Figure 5, one sees a trend as the number of reviews that a person has posted increases; specifically, as the number of reviews increases, the average review tends to decrease, $r(729) = -0.20$, $p < .001$.

This could be explained by a drop-off in the quality of shows – if everyone were seeing the same shows and there were only a small number of genuinely 5-star shows, for example. Given the number of shows involved, however, and how many of these received 5 star ratings from someone, this seems an unlikely explanation. Instead, it seems more likely that we have support for the idea that increased experience in reviewing (and, therefore, seeing more shows) changes the ratings that one is likely to give.

Suggestively, the most prolific reviewers in Figure 5 give average ratings that are more typical of Professional reviewers than the other Amateurs. That is, their mean ratings tend to be between 3 and 4 rather than 4 and 5.

The question remains, however, as to whether this reflects *better* reviews; and the problem is, of course, that as enjoyment of a show is highly subjective, it is possible that what is the *better* (i.e., more predictive) review differs between individuals.

On the basis of these results, for example, one might conclude that the more shows one is inclined to see, then the more similar one's own ratings will be to those of Professional reviewers. If so, then one should weight professional reviews more highly than amateur ones – or, where these are unavailable, downgrade 'overly-enthusiastic' amateur reviews.

Consistency of Different Reviewers

A second consideration in what makes one review *better* than another is their reliability. That is, when two people have seen the same show, are they inclined to give the same rating? A comparison between the Amateurs and

Professionals on such a measure might allow one to have greater or lesser confidence in one group's ratings.

Within the Professional reviewers group, there were 70 shows that had been reviewed by at least 2 reviewers – which yielded a total of 97 pair-wise comparisons (due to some shows being rated by three or four reviewers). Thirty of these had exactly the same rating, with another 40 differing by only half a star. Overall, the average difference between ratings of the same show by professional reviewers was approximately half a star ($M = 0.56$, $SD = 0.52$).

The Amateur group, by comparison, had 228 shows with multiple reviewers, which resulted in 10,401 pair-wise comparisons. This number, however, is dominated by the relatively small number of very popular shows – those on which we see a ceiling effect resulting from the selection bias. The most popular show, for example, has 60 reviews, 58 of which are 5-star – with one 1-star and one 3-star review making up the numbers. This show contributes 1770 unique pair-wise comparisons – over a sixth of the total – and would thus, if included, overwhelm any effects of the inter-rater reliability more generally. To ensure comparability with the Professional results, therefore, only shows that had been reviewed by between 2 and 4 reviewers (the numbers observed in the professional sample) were included in the analyses. This resulted in the removal of 79 shows, leaving 149 and a total of 404 unique pair-wise comparisons.

Of these, 120 had exactly the same rating, 114 differed by half a star and 170 differed by 1 full star or more. The average difference between the amateur reviewers' ratings for these shows was 0.82 stars ($SD = 0.87$), significantly higher than that observed in the Professional reviewers' ratings, $t(499) = 2.83$ $p = .002$.

Discussion

The results paint a complex picture of the relationships between reviewer expertise and the use of aggregation strategies such as the wisdom of crowds for reviews from multiple sources.

Perhaps the single best predictor of show quality (i.e., how much people enjoyed the show) was the total number of reviews that the show had received – reinforcing the assumption that popularity and quality are linked. Note, however, that this is a distinct effect from the wisdom of crowds as the results suggest that we don't need to look at the ratings provided by reviewers at all. Instead, all we need to do is "follow the crowd" and they will lead us to good shows.

In cases without such overwhelming endorsement, however, we are forced to rely on the numerical ratings provided by the expert and amateur reviewers and can run into difficulties in determining what to do.

The first problem we observed in the data was the strong selection bias in the amateur data; because people tend only to pay to see shows that they expect to like, the distribution of star ratings gets shifted to the right – with more 5-star reviews. Added to this is the voluntary nature of amateur reviews, which results in people only writing a review if they are motivated to do so - which, we suggest is most likely when they particularly like or dislike a show. This effect will, therefore, tend to push results even further towards the extremes and, given the effect described above, this will tend to push more people into the very high part of the rating range.

Thus we have a large number of reviews that are relatively uninformative – reflecting the fact that a person predisposed to like a particular show really liked it. A result of this is the lack of discrimination in the amateur data where, because so many reviews give 5-star ratings, it simply doesn't help us to make a decision regarding which of these shows we should attend and short-circuits attempts to use the wisdom of crowds based on median values – as we would end up comparing 5-stars with 5-stars.

A second (but related) concern is that the majority of amateur reviewers (529 of 731) wrote only a single review. Given what we know about people's inability to directly assess values, the use of relative preferences (e.g., converting the ratings to rankings) is a sound method for improving our understanding of what people's expressed preferences actually mean. With only one review per reviewer, however, we cannot meaningfully assess relative preferences.

By comparison, a professional reviewer, while exercising some choice over which shows to see will also have some dictated by their employers and will be asked to write a review of all of the shows that they see. They are, from our data, far more likely to see multiple shows, and have a less-skewed distribution of ratings. They were also, in the subset of shows with a relatively few reviewers, more often in agreement with one another than were the amateurs.

This means that, in relying on professional reviews, one is better able to discriminate between their preferences for those shows that they have seen and also can be more assured that their review is reliable – that is, that another professional reviewer would have a similar opinion.

An addendum to this is that the data support the idea that the difference between amateurs and professional is related to experience/expertise. Amateurs who reviewed larger numbers of shows gave ratings that were more like the professionals. This could suggests that people are, in fact, rating shows on a relative scale but that the single-review amateurs have fewer shows to compare with and thus the chance of the show being amongst the best they have seen is relatively greater. The professionals and high-rate amateurs, by comparison, have a great many shows to compare the current show to and thus the likelihood of it being judged exceptional (5-star) is relatively less.

Caveats

In so subjective a domain, there are, of course, a number of caveats to consider in conjunction with the arguments made above. A primary one, of course, is that we have not made any attempt to look at the types of shows that different people have attended and rated. If we expect that different

people have different tastes in entertainment, then we could conduct a far more fine-toothed analysis of preferences.

This importance of this for the current findings, however, is that one might expect a difference in preferences between professional and amateur reviewers. For example, while purely speculative, it would seem entirely feasible that professional reviewers prefer more serious art whereas the amateurs prefer lighter, comedic events.

If this is the case, then one would have to take into account such between group differences when determining whose reviews should be taken into account when making a decision. That is, knowing that professionals reliably tend to rate a show highly may be of no help at all if it is a type of show that you do not enjoy.

A second caveat is that there has not, as yet, been any attempt to weight or rank the data, which would, as described earlier, be expected to improve the predictive power of ratings – from those reviewers who reviewed multiple shows at least. An appropriate application of such tools, however, requires a fundamental grasp on the nature of the data; a grasp that has been greatly strengthened by the exploratory approach taken here.

Future Research

Given the findings and the caveats noted above, a number of directions for continuing the research suggest themselves. The first is to examine the data in finer detail, dividing shows according to type - to see whether specific reviewers can be identified as having preferences between these.

Data beyond the ratings could also be accessed – for example, using ticket sales to directly measure the popularity of a show rather than simply assuming that number of reviews is a reflection of popularity.

This additional information, used in conjunction with ranking and weighting algorithms, could then be used to generate predictive models for individuals based on the shows that they have seen and how much they enjoyed them and using one half of the data to predict the other – in a similar fashion to the Netflix recommendation algorithms developed as part of the Netflix Prize competition (Bennett & Lanning, 2007).

Finally, experimental work designed to directly measure selection biases in reviews could be conducted, building on the work herein. Similarly, such work could potentially distinguish between alternative judgment strategies – for example, if experts are attempting to provide 'absolute' quality judgments whereas amateurs are just indicated whether they like a show or not.

Conclusions

Within a domain such as entertainment reviews, good decisions can be made by following the crowd – if not always using the wisdom of crowds, per se. Where choices need to be made between shows, however, amateur reviewers ratings tend to cluster too closely around the maximum rating – as a result of selection bias in both show choice and the decision to write a review.

In these cases, therefore, following the advice of more expert reviewers (i.e., professionals and experienced amateurs) seems more likely to provide discrimination as they display less selection bias in their shows seen, meaning that they tend to write reviews of a variety of shows and have clearly discriminable preferences between these.

Acknowledgments

Thanks to ExxonMobil and Santos for their support of the CIBP group in the Australian School of Petroleum; to Michelle Read from the Adelaide Fringe Festival and Simon Evans at BankSA for their assistance in accessing the review databases; and to Dan Navarro, Anna Ma-Wyatt, Steve Begg and three reviewers for their comments.

References

Bennett, J., & Lanning, S. (2007). The Netflix Prize. Proceedings of KDD Cup and Workshop, San Jose, CA.

Galton, F. (1907). Vox populi. *Nature*, 75, 450-451.

Gigerenzer, G., & Todd, P. M. (Eds.). (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2), 231-237.

Malhotra, V., Lee, M. D., & Khurana, A. K. (2005). Domain experts influence decision quality: towards a robust method for their identification. *Journal of Petroleum Science and Engineering, Special Issue.*

Newendorp, P. D., & Schuyler, J. (2000). *Decision Analysis for Petroleum Exploration*. Aurora, CO: Planning Press.

Shanteau, J. (2002). Performance-based assessment of expertise: how to decide if someone is an expert or not. *European J. of Operational Research*, 136(2), 253-263.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, 112(4), 881-911.

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, 15(5), 550-562.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: Random House.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science*, 19(7), 645-647.

Weiss, D. J. (2003). Empirical assessment of expertise. *Human Factors*, 45(1), 104-116.

Welsh, M. B., & Navarro, D. J. (in press). Seeing is believing: priors, trust and base rate neglect. *Organizational Behavior and Human Decision Processes.* Accepted April 6th 2012.

Welsh, M. B., Navarro, D. J., & Begg, S. H. (2011). Number preference, precision and implicit confidence. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1521-1526) Austin, TX: CSS.