

Repeated Judgments in Elicitation Tasks: Efficacy of the MOLE Method

Matthew B. Welsh (matthew.welsh@adelaide.edu.au)

Australian School of Petroleum, North Terrace
University of Adelaide, SA 5005, Australia

Michael D. Lee (mdlee@uci.edu)

Department of Cognitive Sciences, 3151 Social Sciences Plaza A
University of California, Irvine, CA 92697-5100, USA

Steve H. Begg (steve.begg@adelaide.edu.au)

Australian School of Petroleum, North Terrace
University of Adelaide, SA 5005, Australia

Abstract

Recent work has demonstrated the benefits of repeated judgments in improving the accuracy of estimates – both independent and repeated, individual judgments. The More-Or-Less-Elicitation (MOLE) method has previously been shown to improve accuracy and precision of elicitation over single judgment elicitation (Welsh, Lee, & Begg, 2008). In this paper, we show the MOLE method delivers superior gains, in terms of both accuracy and calibration, over repeated-judgment elicitation, while taking significantly less time and not requiring filler tasks to distract participants from previous estimates. We argue that the structure of the MOLE method acts in the same manner as repeated judgment, resulting in multiple searches across the relevant space, without the associated problems of standard repeated judgments methods, such as additional time and participant recall and repetition of prior estimates.

Keywords: Elicitation, Uncertainty, Overconfidence, Repeated judgments.

Introduction

Recent work has drawn attention to the benefits of repeated judgments in estimation/prediction tasks as different as election polling and box office takings projections (Wolfer & Zitzewitz, 2004). That is, having either multiple people all estimate the same value and then combining these (see, e.g., Surowiecki, 2004) or, less intuitively, having the same person make multiple estimates of the same value and combining these (Mozer, Pashler, & Homaei, in press; Vul & Pashler, 2008).

The benefit of this approach is argued to lie in noise reduction. That is, to the extent that the error in one estimate is independent of the error in the others, the average of such estimates will tend to be better than any randomly chosen estimate. Even in the case of a single individual making estimates immediately one after the other, Vul and Pashler (2008) found that the average of two estimates was more accurate than either – supporting the conclusion that the error in each estimate is, at least partly, independent.

Given this finding, it seems possible that incorporating repeated estimation into elicitation techniques, where expert opinions are converted into numerical judgments about the range and probability of possible outcomes (Wolfson, 2001) might yield improvements. Specifically, one would hope to

see improvement in both the accuracy of estimates (how well they correspond to observations from the real world) and people's calibration (how well confidence judgments accord with the proportion of accurate judgments, Lichtenstein, Fischhoff, & Phillips, 1982).

There are, however, problems confronting anyone attempting to use repeated judgments as part of an elicitation technique. Primarily: we don't have the time or resources to contact multiple experts to canvas opinions, which is why a single expert is often relied upon.

On the other hand, asking a single individual to make repeated estimates of the same value may not yield independent estimates because of potential biases such as anchoring and the confirmation bias (Tversky & Kahneman, 1974). That is, people are likely to be strongly influenced by the numbers they have just given or search for reasons that their first estimate is right rather than considering alternatives. Vul and Pashler's (2008) results support this concern, showing that a repeated estimate made immediately after the first, while providing some benefit, is of far less benefit than one made after a three week period.

Similarly, Soll and Klayman (2004) demonstrated that asking a person for two values – one that the true value will be above and one it will be below (with some likelihood in each case) produced better estimates than asking for a single range that the true value was expected to fall within. More recently, Herzog and Hertwig (in press) have shown that individual judgments can be improved via "dialectical bootstrapping" – a process inspired by research into consider-the-opposite debiasing methods (see, e.g., Larrick, 2004) which result in individuals sampling from their knowledge in distinct ways. This method produced significant greater gains in accuracy on date estimation tasks than simply repeating the task but, again, did not equal the benefit gained from including a second person's estimate.

In all of these cases, however, the individual judgments have been limited to just two estimates. That is, the benefit seen is from averaging just two estimates.

To avoid this problem, one needs a way of repeatedly asking an individual about the same estimation task, while avoiding the problems of repetition and confirmation described above. Thus, to be of the greatest benefit, it would seem that an individual's repeated judgments need to be

separated by either a significant period of time or distractor tasks to maximize the independence of the judgments – although, of course, in terms of utility, any significant time delay will cause difficulties. An ideal task would also probe the elicitee’s knowledge repeatedly in different ways so as to get estimates that are as independent as possible.

The MOLE and other methods

The More-Or-Less-Elicitation (MOLE) method (Welsh et al., 2008) seems well suited to offer a way of enabling multiple judgments to be gained from a single individual while avoiding the problems described above. The MOLE relies on repeated, relative judgments (selecting which of two options is thought to be closer to the true value); thereby avoiding the possibility of the elicitee simply repeating their answers. Also, by randomly selecting values from across the parameter space, it ensures that the elicitee is forced to consider new values rather than focusing on their initial value. The MOLE method has been shown to provide superior estimates to basic, single judgment elicitation techniques in terms of both accuracy and calibration (Welsh et al., 2008).

This study, therefore, aims to show whether the benefit results obtained using the MOLE technique is equivalent to the use of other potential methods for obtaining repeated judgments from a single individual - either through direct repetition of the task or repetition with distractor tasks so as to attempt to avoid problems with participants being anchored by or attempting to confirm their earlier estimates repeating values.

Method

Participants

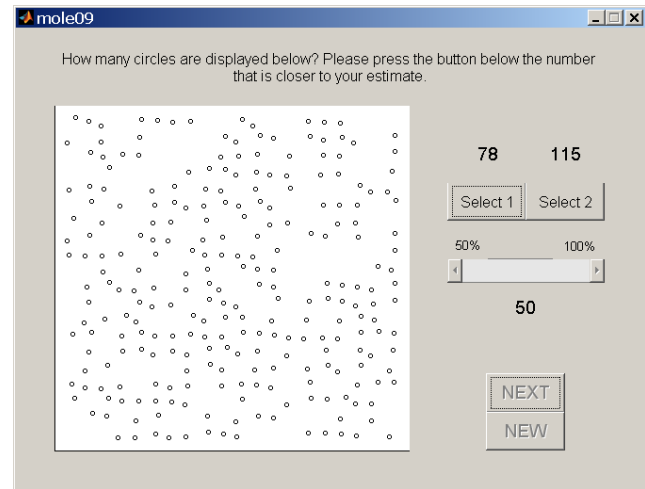
Forty-two participants were recruited; including graduate (12) and undergraduate students (18), university graduates (9) and a small number of non-university educated people (3). Seventeen participants were male and 25 female, with mean age of 28.7 ($SD = 8.9$). Each was given a \$10 book voucher for their participation.

Materials

Three graphical user interfaces (GUIs) were designed along the same lines as those described in Welsh et al (2008), one for each of the three experimental conditions. That is, each, for any given trial, displayed a random array of between 100 and 300 circles but the GUIs differed in terms of the responses available to participants.

Figure 1 shows the MOLE GUI as it appears during a trial – displaying a random array of circles and asking the participant to select which of two numbers they believe is closer to the true number of circles. Once a selection had been made the slider activated and the participant was asked to indicate how confident they were that the number they chose was closer to the true value– from 50% (random or uninformed guess) to 100%.

Figure 1. MOLE GUI.



The other two GUIs, “Repeated” and “Interleaved”, were both variants on the Simple GUI from Welsh et al (2008). The primary difference between these and the MOLE GUI was that, rather than selecting from presented alternatives, participants using these interfaces were asked to enter their estimates into editable text boxes. Specifically participants were asked to give a minimum and a maximum estimate of the number of circles and then rate how confident they were that the true value would fall in that range using a slider.

Procedure

When participants were recruited for the experiment, it was described as an experiment on subjective probability elicitation methods, with no mention made of the experimental hypothesis – to maintain participant naivety.

On arrival, participants read the information sheet and completed a consent form before being seated at a computer for the experiment. A within-subjects design was used, with participants completing all three tasks in a single session in an order determined by a Latin Square.

Participants were allowed a short (2 minute) break between each of the three conditions while the experimenter checked that the data had saved and started the next part of the experiment. Most participants completed the task in less than 40 minutes and none took more than an hour.

Mole Procedure

The MOLE GUI presented a single array of between 100 and 300 circles to a participant, which remained visible for the entirety of the trial. It then randomly drew two alternatives from a range of 0 to 400. The participant was then prompted to select which of these they believed was closer to the true value and indicate how confident they were that this judgment was correct.

If their confidence was 100% - that is, they were 100% convinced that their selection was closer to the true value than the alternative was – then the MOLE GUI used this to truncate the range from which future alternatives would be drawn. For example, if someone was 100% certain that the

true number is closer to 100 than 300, then it was concluded that there was no point offering them values above the midpoint of those options and the MOLE GUI adjusted the *feasible* range accordingly.

Confidence ratings below 100% were not used to adjust the range but rather to assist in the construction of a PDF representing a participant's beliefs about the likelihoods of different numbers as described in the Results section.

Each participant saw a single array of circles in the MOLE condition and was asked to make 10 relative judgments (and confidence ratings) on this single array.

Repeated Procedure

The Repeated GUI also presented a single random array of 100-300 circles that remained visible throughout the trial. Participants were asked to enter a minimum and maximum number representing the range that they thought the true number of circles would fall within.

Participants were also asked to give a confidence rating for how likely it was that the true value would fall within this range (this was done as previous work has indicated that people tend not to use minimum and maximum in absolute terms, see, e.g., Welsh et al., 2008).

Each participant saw only one array of circles in this condition and was asked to give their minimum and maximum value 10 times – having been instructed that we were interested in seeing whether prolonged exposure to the stimulus led them to revise their estimates but that, if it did not, they were free to enter the same numbers on each trial.

Interleaved Procedure

The Interleaved GUI differed from the others in that it presented a series of arrays rather than just one. Specifically, forty arrays of between 100 and 300 circles were presented and the participant asked to give a minimum and maximum number of circles (with confidence rating) for each.

Ten of the 40 arrays, however, were repetitions of a single array – such that participants in this condition completed essentially the same task as during the Repeated condition. These repeated arrays were distributed in a pseudo-random manner throughout 30 distractor trials in order to prevent participants seeing two identical arrays immediately adjacent or noticing any simple pattern (i.e., the experimental arrays were not every fourth trial).

Results

Data Manipulation

Outlier Removal

During analysis of results, discrepancies were observed between a participant's statements regarding their beliefs (made during testing) and the estimates recorded by the GUIs. Specifically, it was observed that the number of circles that participant said they believed most likely was not included within their final range. This was taken to indicate that they had either misunderstood the instructions or had accidentally entered the wrong value. To prevent this and other, unnoticed, errors from impacting results, all

participants' data were analyzed and removed if the error in their estimate on any of the three tasks was identified as an outlier – that is, lying more than 1.5 interquartile ranges above the third quartile (Hodge & Austin, 2004). In all, six participants were identified as having unusually inaccurate estimates in at least one condition and their data were excluded from the subsequent analyses.

PDF Generation

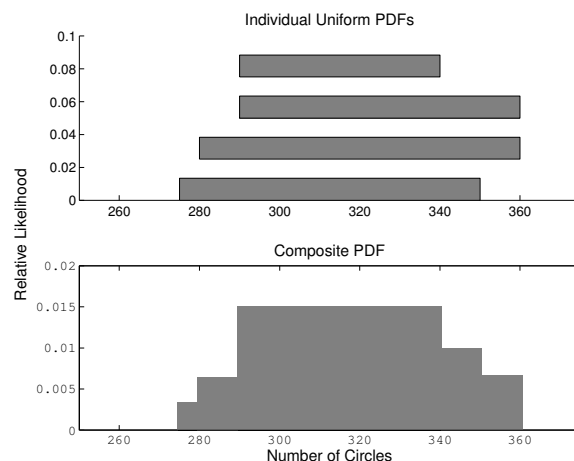
Participants' responses in each of the three conditions were used to construct probability density functions representing their beliefs regarding the number of circles present in the viewed stimuli. The process used to generate the PDFs from the MOLE data was as described in Welsh et al (2008).

That is, for each judgment, if the confidence rating was 100%, the range of possible values was truncated – as described above - and a weight of 0.5 placed across the entire remaining range. For any confidence rating below 100%, by comparison, the range was not adjusted. Instead, weights (determined by the confidence rating) were applied across the entire range. For example, if a person was offered the options 200 and 300 and selected 200 with 75% confidence, a weight of 0.75 was applied to all points within the current range closer to the selected option (200) and a 0.25 weight to all other locations.

This process repeated for each of 10 judgments made by an individual on a single array and a PDF built up in this way. After the 10th judgment was incorporated, the PDF was corrected by removing all weight lying outside the final range and then subtracting 99% of the lowest weight (these corrections dealt with excess weight added during early trials when much of the range is still considered feasible).

Finally, the Beta-distribution that minimized summed squared differences from the resultant PDF was calculated and this used to represent the participant's beliefs regarding the likelihood of various numbers.

Figure 2. Example of PDF construction from individual, uniform PDFs.



In the Repeated and Interleaved conditions, by comparison, a somewhat simpler (although clearly related) method of PDF construction was used. Each participant

estimated 10 ranges (Minimum to Maximum) for a given stimulus and each Min-Max range was used to define a uniform PDF to represent their beliefs regarding the number of circles displayed for each instance. The participant's overall, or composite, PDF was then taken to be simply the average of the 10 uniform PDFs.

For example, the uniform PDFs in the top subplot of Figure 2 are the actual ranges given by a participant in the first four stages of the Repeated condition and the composite PDF is simply the sum of the individual heights at each value divided by the number of PDFs being considered (4 in this figure but 10 in the experiment).

Composite PDFs were then used to calculate the participant's estimates of the mean and the range for determination of accuracy and calibration. For example, the composite PDF in Figure 1 yields a final range of 275-360 and a mean estimate of 318.1.

Comparison of Elicitation Methods

To compare elicitation methods a number of measures are required - to assess the accuracy of estimates and the adequacy of estimated ranges. For accuracy, correlations between the true and estimated number of circles were calculated, along with absolute percentage error. Calibration, on the other hand, was examined by comparing the proportion of ranges that contained the true value (hits) and the confidence statements made by participants.

Table 1. Summary of elicitation technique performance.

Technique	Accuracy		Calibration	
	r	!% Error!	% Hits	Confidence
Single: R*	0.42 (0.05)	31.9 (1.7)	46.1 (4.6)	74.3 (3.1)
Single: I*	0.39 (0.14)	27.8 (4.3)	56.9 (4.2)	73.1 (1.3)
Repeated	0.44	31.3 (22.9)	69.4	100
Interleaved	0.49	23.5 (20.1)	88.9	100
MOLE	0.66	22.4 (15.8)	91.2	100

* - Single refers to data from individual trials of the repeated (R) and Interleaved (I) conditions. The values in these rows are thus means and SDs from the 10 trials. Means and SDs in the remaining rows are calculated from the 36 participants.

Table 1 summarizes these key statistics used to look for differences between the elicitation techniques. The columns hold results for five elicitation methods - the three described above and summarized results from the individual trials of the repeated and interleaved conditions - representing single judgment elicitations. These results are discussed in greater detail in the following sections.

Repeated versus Single Judgments

The first comparison that needs to be made is between the repeated judgments elicitation techniques and their single

judgment equivalents. Looking at Table 1, one sees the mean correlation between the true and estimated number of objects across the 10 trials of the Repeated method was 0.42, while the composite estimate correlates slightly better at 0.44. Similarly, the correlation between the true value and composite value in the Interleaved condition, at 0.49, is somewhat stronger than the average of the correlations from the 10 Interleaved trials, at 0.39. In both cases, however, the composite estimates' correlation is within a single standard deviation of the mean value of the individual correlations, $z = 0.40$ and 0.71 respectively, so concluding that the repeated judgments resulted in improvement is a long bow to draw.

A similar pattern can be seen in the error scores, with no improvement from the use of the Repeated method over the individual trials comprising it and a small (and again statistically insignificant, $z = 1.0$) improvement from the use of repeated judgments in the Interleaved conditions.

In the calibration data, however, we see a clear change resulting from the use of repeated judgments for both the Repeated and Interleaved conditions. For the former, the percentage of hits increases by 23.3%, $z = 5.1$, while for the latter it increase by 32.0%, $z = 7.6$. It is harder to say exactly what this means in terms of calibration, of course, as the confidence ratings given by participants apply only to individual judgments. But, assuming a 100% confidence rating for the composite judgments, there seems to be little difference in calibration using the Repeated method - 28.2% overconfidence in the individual judgments compared to 30.6% in the composite, where overconfidence is defined as the difference between the percentage of hits and the confidence level. Comparing the composite and individual judgments from the Interleaved method, however, one sees a small decrease in overconfidence from 16.2% to 11.1%.

Overall, there seems to be weak evidence that the use of repeated measures is of benefit to the accuracy and calibration of elicited ranges but only where the repeated judgments are interspersed amongst distractor tasks.

Accuracy

Turning now to the primary comparison between the three elicitation methods, Figure 3 shows scatterplots between estimates made in each condition and the true value.

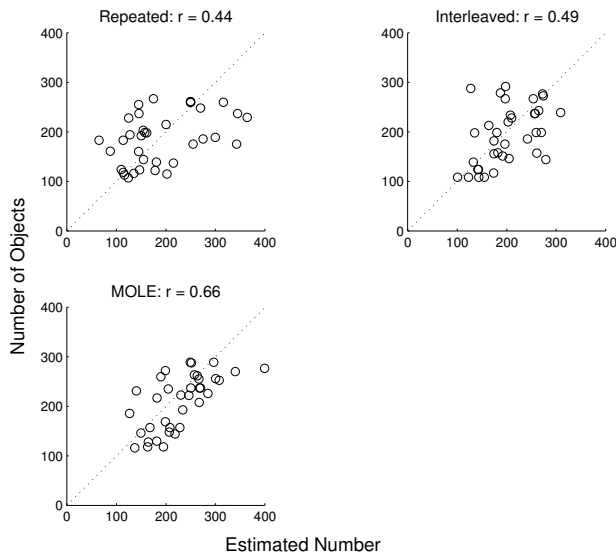
Looking at Figure 3, one can see that estimates made under all three conditions show evidence of some degree of accuracy - with a positive relationship observed between the estimates and the true value. The strength of the relationship, however, varies from 0.44 in the Repeated condition to 0.66 in the MOLE. All of these correlations are significant at the .01 level and the MOLE results are significant at $p < .001$ indicating that estimates elicited using the MOLE are the best predictors of the true value.

A correlational study, however, while indicating the strength and direction of a relationship misses a key factor in determining accuracy - the fit between the ideal and the observed data, represented in Figure 3 by the dotted line.

Looking at the results in column 2 of Table 1, one sees the percentage error scores achieved by participants in each elicitation method. Once again, the MOLE technique is the

most accurate, with a mean error of 22.4%. The Interleaved method does almost as well, with a mean error of 23.5%, while the Repeated is, again, the worst with a mean error of 31.3%. A repeated-measures one-way ANOVA conducted comparing these results, found a near significant result, $F(2,70) = 2.41, p = .097$. Paired sample t-tests confirmed that, considered separately, both the MOLE and Interleaved methods produced better results than the Repeated, $t(35) = 1.81$ and $1.70, p = .040$ and $.049$, respectively.

Figure 3. Scatterplots of true and estimated number of circles in arrays. $N = 36$ in all cases.



Calibration

The second measure of the adequacy of an elicitation technique is the level of calibration achieved by people using it. That is, how often the ranges elicited from them contain the true value compared to how often the participant’s confidence level indicates they should.

In all three conditions participants made confidence judgments after every individual judgment (selection between options or estimation of range). These confidence ratings, however, do not directly relate to the overall confidence that the true value will fall within the final range calculated from a participant’s PDF. Instead, as was done with the MOLE results in Welsh et al (2008), the final range is treated as a 100% confidence interval when calculating overconfidence for each technique. The calibration data for the three techniques is shown in Table 2.

Looking at the data in Table 2, one sees that the MOLE condition produced the best calibrated results, with 91.2% of the composite ranges containing the true value (c.f. 90.6% in Welsh et al (2008)). By comparison, the Interleaved condition ranges contained the true value 88.9% of the time and the Repeated condition 69.4%.

Clearly, the percentage of hits observed depends on two things – the accuracy and of the range. The accuracy of judgments was discussed above so here a repeated-measures one-way ANOVA was run to determine whether the difference in range width was significant across conditions.

This indicated a significant effect, $F(2, 70) = 18.7, p < .001$. Paired sample t-tests revealed that this significant result was caused by differences between the Repeated condition and the other two, $t(35) = 6.63$ and $4.19, p < .001$ for the Interleaved and MOLE respectively, with difference between range widths in the Interleaved and the MOLE conditions approaching significance, $t(35) = 1.45, p = .077$.

Table 2. Calibration data for elicitation techniques including mean and SD range widths.

Technique	% Hits	Conf.	Overcon.	Range Width
Repeated	69.4	100%	30.6	167.5 (138.8)
Interleaved	88.9	100%	11.1	332.6 (157.1)
MOLE	91.2	100%	8.8	289.4 (127.5)

Note: ‘Conf’ refers to the assumed confidence rating of 100% for composite ranges. ‘Overcon.’ is the difference between the confidence and the number of hits.

Time

A third important consideration for any elicitation technique is its ease of use. For the purposes of this study, we regard the time taken to complete a single elicitation to be one measure of this. Table 2 shows the time taken to complete an elicitation task under each of the three conditions.

Looking at the data in Table 3, it seems clear that the MOLE is easily the fastest of the techniques, taking an average of just 3 minutes to complete. The Repeated method also fares relatively well, taking between 4 and 5 minutes to complete while the Interleaved method required an average of more than 17 minutes to complete. Of course, this is not surprising given that the Interleaved condition required four times as many judgments to be made as the Repeated – thereby ending up four times as long.

Table 2. Time to complete task by condition

Condition	Mean Time (secs)	SD
Repeated	252	87
Interleaved	1033	377
MOLE	180	90

A repeated measures ANOVA confirmed the significance of the differences in time taken, $F(2, 70) = 194.8, p < .001$, and paired sample t-tests indicated that all three conditions differed significantly from one another, $t(35) = 13.5, 6.1$ and 14.6 , for the R vs I, R vs M and I vs M comparisons respectively, $p < .001$ in each case.

Discussion

The results presented above offer some support for the use of repeated judgments in elicitation tasks – in line with expectations. The Repeated method, subject to all of the standard problems with repeated individual judgments, unsurprisingly, failed to improve either the accuracy or calibration of elicited ranges. By contrast, there is evidence that the Interleaved method, which aimed to avoid these

problems by locating the experimental trials within a series of distractor tasks, yields a benefit. Specifically, there was a small increase in the accuracy of estimates but also a significant increase in the width of elicited ranges and a commensurate decrease in overconfidence. Finally, the MOLE method was clearly superior to either of the other methods – being more accurate, generating less overconfident ranges and taking the least time to complete.

The question as to why this might be is an interesting one. Clearly, the MOLE technique forces people to consider options that they otherwise may not have resulting in multiple searches of their beliefs about the stimulus. That the Interleaved condition does not provide equal benefit, however, indicates that there may be more to it than this.

One possibility is that the use of relative, rather than absolute, judgments in the MOLE allows people to use more finely tuned cognitive abilities – taking advantage of the fact that people tend to be better at making relative rather than absolute judgments (see, e.g., Gigerenzer & Selten, 2001).

Limitations

Despite the strength of the results, there are limitations that should be addressed. First, there is a question of whether people in the Interleaved condition realized that one stimulus was repeating. One participant did state they believed this to be the case but the much wider ranges in the Interleaved condition compared to the Repeated argues against this having been a common feeling.

Secondly, the MOLE technique has an advantage over the others in that it starts with a pre-specified range from which to draw its numbers and thus limits participant's estimates to this 0-400 range, whereas people in the other conditions were free to make any estimate they chose. While this is true, it should be noted that none of the estimates from the Repeated and Interleaved conditions actually fell outside the 0-400 range once outliers had been removed.

Finally, it should be noted that the overconfidence scores used herein are based on assumptions regarding the confidence that a participant would place in the final composite range from any of the conditions rather than direct measurement thereof.

Future Directions

Given the promising results, further development of the MOLE to address some of the concerns raised above and improved its ability to capture people's beliefs need to be considered. For example, starting with a wider pre-generated range to avoid the above criticism or utilizing a person's partially constructed PDF during testing to better guide the values that are presented.

Another improvement would be to add in, following the MOLE technique as it stands, an evaluation stage where participants are shown the range/PDF they have generated and are asked to indicate how likely it is that the true value will fall within this region.

Conclusions

In all, the results support the use of repeated individual judgments in elicitation tasks but only under circumstances

where the standard problems with this process can be overcome – either through the use of time delays between judgments or other means such as distractor tasks.

Where these are not possible, however, the MOLE seems to yield the benefits of repeated measures without these problems – its technique of asking for repeated, relative judgments avoiding the problems of simple repetition without the need for lengthy delays or complex experimental design.

Acknowledgments

MBW and SHB are supported by ExxonMobil and Santos through their support of the CIBP at the Australian School of Petroleum. The authors wish to thank Mark Steyvers and Danny Oppenheimer for their suggestions.

References

- Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded Rationality: the adaptive toolbox* (Vol. 84). Cambridge: Massachusetts: MIT Press.
- Herzog, S. M., & Hertwig, R. (in press). *Psychological Science*, Accepted 15th July 2008.
- Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 85-126.
- Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 316-337). Malden, MA: Blackwell.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Mozer, M. C., Pashler, H., & Homaei, H. (in press). Optimal predictions in Everyday Cognition: The Wisdom of Individuals or Crowds. *Cognitive Science*, Accepted July 14th 2008.
- Soll, J. B., & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(2), 299-314.
- Surowiecki, J. (2004). *The Wisdom of Crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York, NY: Doubleday.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19, 645-647.
- Welsh, M. B., Lee, M. D., & Begg, S. H. (2008). More-Or-Less Elicitation (MOLE): Testing A Heuristic Elicitation Method. In V. Sloutsky, B. Love & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 493-498). Austin TX: Cognitive Science Society.
- Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18, 107-126.
- Wolfson, L. J. (2001). Elicitation of probabilities and probability distributions. In E. Science (Ed.), *International Encyclopedia of the Social Sciences* (pp. 4413-4417): Elsevier Science.