

Running head: MORE-OR-LESS ELICITATION (MOLE)

More-Or-Less Elicitation (MOLE): Incorporating an  
Individual's Repeated Judgments into Elicitation.

Matthew B. Welsh

University of Adelaide

Australia

Michael D. Lee

University of California, Irvine

USA

Steve H. Begg

University of Adelaide

Australia

### Abstract

Elicitation of people's opinions is a central methodological challenge for psychology, with impacts in technical and industrial settings. Simply asking for a point estimate or range, however, is subject to biases such as overconfidence and, thus, methods that avoid these are required. We present a novel elicitation method, More-Or-Less Elicitation (MOLE) which, rather than requiring people make single, absolute judgments, asks for repeated, relative judgments. MOLE uses these sets of relative judgments to construct PDFs representing a person's beliefs. Study 1 compares these PDFs with traditionally elicited ranges; concluding that MOLE improves both accuracy and precision of elicited ranges. In Study 2, we show MOLE delivers superior results to alternate repeated-judgment methods, while taking less time and not requiring distractor tasks. We argue MOLE results in multiple searches within participants' 'belief space', without the problems normally associated with repeated inquiry from a single individual – such as participant recall and repetition of prior estimates.

**Keywords:** Elicitation, Uncertainty, Overconfidence, Repeated judgments.

## More-Or-Less Elicitation (MOLE): Incorporating Repeated Judgments from Individuals

Many technical disciplines share with psychological research the problem of eliciting information from people; that is, translating peoples' beliefs into useable data (Wolfson, 2001). Of particular interest is how to best achieve this under uncertainty, where there is no single, correct answer but rather the "correct" response for an individual to make will vary according to their own level of knowledge about the topic (Morgan & Henrion, 1990).

The reason so much interest is vested in this area is that, despite elicitation's ubiquity, argument continues about the best way to elicit information and people are still subject to many biases, so that elicited responses are less accurate than elicitors would wish (see, e.g., Hawkins, Coopersmith, & Cunningham, 2002; Lichtenstein, Fischhoff, & Phillips, 1982; Welsh, Begg, & Bratvold, 2007).

This paper discusses elicitation and some of its problems. It then proposes, as a possible solution, the use of *heuristic*-based methods – that is, elicitation methods based on simple judgments, such as which of two options is closer to the true value (i.e., the value the participant believes the stimulus takes). Such a method, called More-Or-Less Elicitation (MOLE), is described – in which people are asked to make repeated, *relative* judgments rather than to make single, absolute judgments. The MOLE method is then tested against both single and repeated judgments elicitation methods.

### *Elicitation*

The elicitation of uncertainty is the conversion of individual or group's beliefs into numerical form – whether a point estimate or a probability distribution. Generally, this is done not for its own sake but to, for example, predict future outcome ranges, or provide inputs for forecasting models (Morgan & Keith, 1995).

In order to be of benefit, elicited values need to be accurate. Accuracy, in this case, however, refers to two separate ideas. The first sense, which we might call objective accuracy, is the one that

naturally springs to mind: elicited values need to accurately reflect the probability of an event occurring. Equally important, however, is subjective accuracy: how well elicited values map onto an elicitee's beliefs. The problem for elicitors is that the two are not easily separated. Instead we have to rely on relatively crude measures like overconfidence/calibration scores (Lichtenstein et al., 1982), which primarily measure objective accuracy even though, from the elicitor's point of view, a measure of subjective accuracy is being sought.

### *Problems for Elicitation*

Standard findings in the elicitation literature are that people's best guesses are anchored by previously seen values and that they are overconfident, producing too narrow ranges of possible outcomes (Tversky & Kahneman, 1974). There is also evidence, however, that this is not entirely due to inaccuracies in people's beliefs. Specifically, different elicitation techniques result in different responses; Winman, Hansson and Juslin (2004), for example, demonstrate that having people evaluate a range rather than produce one leads to less overconfidence in their responses.

There are also concerns about the effect of question order within an elicitation task. These date back to Tversky and Kahneman's (1974) paper, where they suggested that anchoring on an initial best guess might be a cause of overconfidence. Research into this idea, however, has been mixed with, for example, Russo and Schoemaker (1992) finding the predicted effect but Block and Harper (1991) and Juslin, Wennerholm and Olsson (1999) finding the opposite. To complicate matters further, there are concerns regarding the level of control over question order in some of these studies. For example, Block and Harper (1991) used answer booklets which, while having questions in a set order, could not insure they were answered in that order.

### *Debiasing Elicitation*

Given these problems, significant work has gone into attempts to debias elicited values. Early work (summarized in Morgan & Henrion, 1990), however, indicated little success in reducing overconfidence and none for anchoring.

As noted above, however, there are some techniques known to reduce overconfidence, including Winman et al.'s (2004) use of interval assessment, and the use of long-term repeated feedback (Lichtenstein et al., 1982). Additionally, the use - by expert elicitors - of counterintuitive examples (lying outside the initial range) has been shown to be effective in reducing overconfidence (Hawkins et al., 2002). This remedy, though, requires an expert elicitor to be on hand to ask the right sorts of questions and leaves open the question of whether simply drawing people's attention to regions of the possibility space outside their initial range is helpful in the absence of expertise.

Regardless, none of these techniques eliminates overconfidence – excepting specific cases such as weather forecasting, where repeated feedback seems to have resulted in good calibration (Murphy & Winkler, 1977).

*Repeated Judgments in Elicitation.* More recently, research has drawn attention to the benefits of repeated judgments in estimation/prediction tasks as different as election polling and box office takings projections (Wolfers & 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 (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 two estimates immediately one after the other, Vul and Pashler (2008) found that the average of these 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 (how well confidence judgments accord with the proportion of accurate judgments, Lichtenstein et al., 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 (2009) 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 Method*

Given the problems with elicitation and the observation that question format has a large impact on the elicited responses, it is worth considering more radical departures from the standard elicitation methods. For example, the work of Gigerenzer and others (Gigerenzer & Selten, 2001; Gigerenzer & Todd, 1999) on bounded rationality has yielded insights into the sorts of questions that the human mind seems most comfortable working with.

One observation is that people are better at making relative judgments than absolute ones (Gigerenzer & Selten, 2001). This is consistent with Winman et al.’s (2004) observation that people are better at evaluating than generating ranges. Combining this insight with the observation that counter-intuitive examples can reduce overconfidence (Hawkins et al., 2002) and the observations above regarding the benefits of repeated judgments leads to the idea that asking a series of questions - covering the range of possibility - rather than allowing a person to hone in on a small region of outcomes, thereby excluding other possibilities, may yield better results.

The idea of such an elicitation method – using relative judgments – was first explored in Welsh, Begg, Bratvold and Lee (2004). This found a benefit but relied heavily on assumptions about the underlying distribution required to create a probability distribution from the relative judgments. The current goal was, thus, to revise this method in such a way as to create an elicitation method that makes minimal assumptions in a principled manner to produce elicited responses that are less subject to overconfidence than alternative methods requiring direct estimation of values.

Such a method also has the potential to allow repeated accessing of a single person’s beliefs in a manner that avoids obvious problems with repetition and anchoring as, at no point, is the elicitee

required to actually state an estimates.

### Experiment 1

The first experiment was designed to compare estimates elicited using the MOLE method with those achieved using traditional range estimation methods; the goal being to determine whether the use of a series of relative judgments (rather than a single absolute judgment) would result in superior estimates – in terms of both the accuracy and calibration of responses.

Three elicitation methods were chosen for comparison with the MOLE: a simple range estimation task, where participants gave a minimum and maximum estimate; a triangular estimation task, to test the impact of providing a best guess prior to estimating the range; and an iterative elicitation task, to assess the impact of calling participants' attention to regions outside their initially estimated range. Both of these variants of the simple range elicitation method were selected as evidence exists that these might provide some benefit in terms of reducing overconfidence.

### *Method*

#### *Participants*

Participants were 40 undergraduate students from the University of Adelaide. Four, however, were excluded due to computer errors during testing leaving 36 (10 male and 26 female) with a mean age of 20.1 ( $SD = 1.9$ ). Participants received a \$10 book voucher for their participation.

#### *Materials*

Four graphical user interfaces (GUIs) were developed to enable automated testing of participants using each of the elicitation methods chosen for examination. All of the GUIs displayed an array of circles, from 100 to 300 (determined randomly at each trial) and elicited the participant's beliefs regarding the number of circles - in accordance with the varying elicitation techniques.

For each of the elicitation techniques, the same basic GUI layout was used, with only the questions being asked and the buttons that could be used to respond being different. For example, Figure 1 shows the layout as seen during More-Or-Less Elicitation (MOLE) condition, asking

participants to select which of two values is closer to their estimate. The GUI controls were sequentially locked and unlocked to ensure that participants answered each question before continuing to the next. This ensured that participants completed the questions in the prescribed order.

### *Procedure*

Participants were tested on four elicitation methods, described below. Participants, over the course of an hour, completed 10 trials under each condition after being sorted at random into four groups to allow counterbalancing for possible order/learning effects as shown in Table 1.

*Simple Elicitation.* In this condition, participants were asked to provide a minimum and maximum value for the number of circles. Following this, they indicated how confident they were that their range contained the true value. This was done using a slider similar to the one seen in Figure 1 but capable of taking any integer value from 0 to 100%<sup>1</sup>.

*Triangular Elicitation.* In this condition, participants were asked to provide a best guess prior to giving their minimum and maximum values – thereby providing sufficient information to produce a triangular distribution. Again, after making estimates, they were asked to indicate their confidence on a 0-100% scale.

*Iterative Elicitation.* In this condition, participants were asked to provide an initial range as in the Simple Elicitation condition but then shown values for the minimum and maximum that lay outside their own range - which were described as having been elicited from “previous participants” but which actually were always calculated by the program to lie outside the initial range (60% of the initial minimum and 140% of the initial maximum). Participants were then given the chance to adjust their estimates of the minimum and maximum. Once happy with their estimates, they were asked to indicate their level of confidence that the true value would fall inside their range on a 0 to 100% range.

*More-Or-Less Elicitation.* In the MOLE condition, participants did not directly estimate values. Rather, they selected which alternative from a pair of values (randomly generated from a

range from 0 to 400) was closer to their estimate. After each choice, participants were asked to indicate their confidence that their selection was actually closer to the true value than the alternative - on a 50% (guessing) to 100% (certain) range.

This process was repeated 10 times during each trial and the final range of feasible values recorded (i.e., those the participant's answers did not rule out). Additionally, the confidence ratings were used to create a subjective PDF as described below.

Whenever a confidence rating of 100% was given, any values lying closer to the unchosen value were excluded from the experimental range and then weight of 0.5 was added uniformly across the remaining range. This was seen as a logical consequence of someone being 100% confident that the true value was closer to their selection than the alternative – that being that no value closer to the alternative was considered possible by the participant. If, however, the confidence level was less than 100%, weight was added to each end of the range separately according to the level of confidence.

Figure 2 shows how two stages of this process might progress, starting with a range of possible value from 90 to 150. In the top half of the figure, the person has been shown two values: 135 and 150 (highlighted) and stated with 100% confidence that 135 is closer to the true value. This means that values above 142.5 (the midway point of 135 and 150) will no longer be considered. An equal weight of 0.5 is then applied across the entire remaining range, indicating ignorance about where in that range the person believes the true value lies.

In the lower half of Figure 2, the person is then shown the values of 105 and 130 and states that they are 75% confident that 105 is closer to the true value. This results in a weight of 0.75 being applied from the current minimum up to the midpoint of the two values (117.5) and a weight of 0.25 from the midpoint up to the current maximum – reflecting the fact that the person's confidence statement indicates that they believe a value closer to the lower option is three times as likely as one closer to the high option.

In this way, over the course of a trial, a PDF was built up. At the end of each trial, this PDF was corrected by removing all weight from areas outside the final feasible range and then adjusted by subtracting 99% of the lowest weight from all remaining areas. Finally, the Beta–distribution that minimized summed squared differences from the resultant PDF was calculated.

### *Results*

As described above, while overconfidence is generally used as the primary measure of the efficacy of an elicitation method of the sorts used herein, this can be further divided into the accuracy and the precision of the elicited responses. Results relating to the primary hypothesis (that repeated, relative judgments would result in superior estimates than traditional elicitation) are therefore described below in terms of all three concepts: overall overconfidence, precision and accuracy.

#### *Overconfidence*

Overconfidence, in terms of elicited ranges, is measured from the degree of ‘coverage’ achieved (i.e, how often the elicited range was correct - that is, contained the true value) and participants’ stated levels of confidence. Table 2 shows this data for each of the four conditions.

It is clear from Table 2 that all three techniques requiring participants to estimate absolute ranges resulted in less than 30% hits, despite the stated confidence of the participants averaging more than 70%. By comparison, the MOLE, with its assumed 100% confidence level, resulted in 90.6% coverage. (The confidence level is ‘assumed’ as participants in the MOLE condition did not directly rate the likelihood of the true value falling within their final range, rather it was assumed that their final range contained all of the values they considered feasible.)

To determine whether the differences between the methods were statistically significant, a repeated measures ANOVA was conducted. The first indicated that there were significant differences between the number of hits achieved by participants under the four conditions,  $F(3, 83) = 123.8, p < .001$ . Paired sample t-tests with Bonferroni corrections were used for each unique pair of elicitation methods to determine which conditions differed from the others and these indicated that only the

MOLE condition differed significantly,  $t(35) = 14.1, 15.7$  and  $12.4$  (from the Simple, Triangular and Iterative, respectively),  $p < .001$  in all cases.

The question remained, however, as to whether the improvement in calibration in the MOLE data resulted from an improvement in precision, accuracy, or both.

*Precision.* Precision reflects the subjective aspect of accuracy (i.e., how well a response matches the person's knowledge and beliefs). The primary measure of precision in an elicited range is its width. Figure 3 shows the mean range width under each of the methods described above, along with the variability, as measured by the standard deviation.

Looking at Figure 3, it seems clear that an improvement in the appropriateness of participants' levels of precision plays a significant role in the observed reduction in overconfidence. Specifically, participants' responses to the MOLE technique are far less precise, giving much wider ranges on average than in any of the other three conditions.

The implication of this is that participants in the other conditions were *too* precise. That is, their ranges were far narrower than their level of knowledge warranted. A repeated measures ANOVA, run using the participants' mean ranges in each condition, confirmed that the difference was highly significant,  $F(2, 59) = 75.9, p < .001$ . Once again, paired sample t-tests with Bonferroni corrections were used post-hoc confirming that only the MOLE results differed from the other conditions,  $t(35) = 8.6, 12.1$  and  $9.4$  for comparisons with the Simple, Triangular and Iterative methods, respectively,  $p < .001$  in all cases.

*Accuracy.* To assess the objective accuracy of participants' responses under each elicitation condition, the mode of each elicited range was compared with the true value. For the Simple and Iterative elicitation conditions, a uniform distribution was assumed and thus the mean (midpoint) was substituted for the mode. For the Triangular, the mode was the "most likely" value given by the participant. Finally, for the MOLE, the mode was calculated using the  $\alpha$  and  $\beta$  values calculated

from the beta distribution most closely fitting a participant's subjective PDF, giving  $M = (\alpha-1) / (\alpha+\beta-2)$ . Scatterplots showing these data are shown as Figure 4.

Figure 4 suggests that only in the MOLE condition did participant estimates accurately track the number of objects in the stimuli. The correlation between the means of the estimated range and true values was moderately high and highly significant,  $r(338) = 0.64, p < .001$ , whereas correlations between the true values and the remaining elicited means were all extremely low,  $r(338) = -0.01, -0.10$  and  $-0.02$  for the Simple, Triangular and Iterative method respectively,  $p > .05$  in all cases.

### *Other Findings*

*Best Guesses and Overconfidence.* One of our secondary research goals was to determine whether requiring participants to give a best guess prior to fixing their confidence interval's endpoints would affect its width and thus their levels of overconfidence – as previous research on this question has been mixed.

Looking at the data in Table 2 and Figure 3, one sees little difference between the ranges provided in the two conditions of interest (Simple and Triangular). While participants in the Triangular condition gave, on average, narrower ranges ( $M = 84.7, SD = 61.8$ ) than they did in the Simple condition ( $M = 100.3, SD = 105.0$ ) the confidence intervals in Figure 3 indicate no significant difference between these values. A repeated measures ANOVA, similarly, compared the mean levels of confidence indicated by participants in the three conditions where this was directly assessed (i.e., all but the MOLE) and this found no significant differences between the conditions,  $F(2,67) = 2.26, p = .112$ . Even were the differences significant, however, overconfidence would not be greatly affected as the 3.3% decrease in the number of hits is offset by the 2.9% decrease in stated levels of confidence.

*Iterative Elicitation.* The final research question related to whether participants could be prompted to reconsider and widen their ranges by providing them with reasons to reconsider values outside their initially estimated range. Looking again at Table 2 and Figure 3, one sees that there seems to be a weak effect in line with expectations. Participants' ranges in the Iterative condition

were wider than in the Simple condition ( $M = 105.5$ ,  $SD = 85.0$  compared to  $M = 100.3$ ,  $SD = 105.0$ ) but not significantly so as examination of the CIs in Figure 3 shows. Similarly, the difference in confidence, while noticeable in Table 2, is not significant – as the repeated measures ANOVA described above indicated.

### *Discussion*

Our results show a clear benefit to the use of the MOLE technique in terms of both the precision and the accuracy of elicited ranges. We found little support, however, for the role of initial best guesses or simplistic emphasis on counter-intuitive values in improving elicitations. These results are discussed in greater depth below.

### *Other Research Questions*

While previous work has found estimated ranges to be either narrowed (Russo & Schoemaker, 1992) or widened (Block & Harper, 1991) by initial best guesses, the present study, despite stronger control over question order, found no clear effect - although a weak trend was seen towards narrowed ranges resulting from initial best guesses. As such, this remains an open research question, with further work required to tease out the intricacies of this variable effect.

Neither was any evidence found that simply indicating to people that other people had made estimates well outside their own range had any impact on revisions of those estimates. It could, however, be that participants realized that these “other participants” were computer generated. As such, future research will determine what sort and how much counter-intuitive evidence people need to provoke them into changing their mind – or whether this only occurs with the presence of a known expert (Hawkins et al., 2002).

### *Limitations*

There are, however, some caveats regarding the MOLE method as used in Experiment 1. Firstly, the MOLE required participants to spend more time observing the stimulus and thus some of the effect may simply be noise reduction – although this would seem only to explain improvements

in accuracy, not precision. This also has the effect of increasing the effort required per trial with the resultant problem that more participants gave nonsensical answers indicative of random button pushing in the MOLE (18 of the 20 excluded trials). This is, however, unlikely to cause a problem in applied setting as large numbers of values tend not to be elicited simultaneously.

Additionally, given the fact that the stimulus display set out its circles in rows and columns the additional time could, potentially, have allowed participants to more accurately gauge or even count the circles – although no evidence of this seen during testing.

## Experiment 2

The results of Experiment 1 strongly supported the idea that the MOLE is a superior elicitation method to traditional range estimation. There were however, questions arising out of the results that still required examination – specifically as regards the repeated judgments aspects of the task.

The MOLE method seems well suited to offer a way of enabling multiple judgments to be gained from a single individual while avoiding the problems described above. As the MOLE relies on repeated, relative judgments (selecting which of two options is thought to be closer to the true value) it thereby avoids 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 question of how much of this benefit could be achieved using other repeated judgment methods, however, needs to be answered in order to determine whether it is just the repetition or the combination of repetition and relative judgments that is providing the benefit. It is also necessary to confirm that the benefit off the MOLE is not simply the result of the additional time spent by participants examining the stimulus figure – as suggested above.

This study, therefore, aimed to show whether the benefit resulting from 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. These tasks, necessarily, took as long or longer than the MOLE to complete and, as such, were also expected to indicate whether the superiority of the MOLE resulted from noise reduction due to increase time spent on each elicitation task.

### *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 above - one for each of 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 5 shows the MOLE GUI as it appeared 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. Note that in all Experiment 2 GUIs, jitter was added to the location of circles within the display so as to prevent the circles forming rows and columns.

The other two GUIs, “Repeated” and “Interleaved”, were both variants on the Simple described above. 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*

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. Only a single trial was conducted under each condition and most participants completed the task in less than 40 minutes and none took more than an hour.

*Mole Procedure.* The MOLE GUI worked exactly as described above with one exception; that being, participants here completed only a single trial rather than the 10 completed in Experiment 1.

*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.

While each participant saw only one array of circles in this condition they were 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). By interleaving the experimental trials amongst distractor trials, it was expected that some of the problems with using repeated judgments could be overcome.

## *Results*

### *Data Manipulation*

*Outlier Removal.* During analysis of results, discrepancies were observed between a participant's statements regarding his 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 exactly as described above.

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 6 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 3 summarizes these key statistics used to look for differences between the elicitation techniques. The rows 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 3, 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%.

Thus, 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 – as would be expected given the argument from the independence of error.

*Accuracy.* Turning now to the primary comparison between the three elicitation methods, Figure 7 shows scatterplots between estimates made in each condition and the true value.

Looking at Figure 7, 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 7 by the dotted line.

Looking at the results in column 2 of Table 3, 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.

*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 Experiment 1, 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 4.

Looking at Table 4, 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 Experiment 1). 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,  $t(35) = 1.45, p = .077$ .

*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 5 shows the time taken to complete an elicitation task under each of the three conditions.

Looking at the data in Table 5, 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.

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. 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 observation that the MOLE produces the best results while taking the least time to complete also undermines the suggestion raised following Experiment 1, that the advantage of the MOLE over the traditional range elicitation techniques resulted from noise reduction due to participants spending longer looking at the stimulus.

### *Limitations*

Despite the strength of the results, there is a limitation that should be addressed. Specifically, there is a question of whether people in the Interleaved condition realized that one stimulus was repeating. If this was the case, then the potential benefit of repeated judgments would be reduced by the same effects restricting the benefits in the Repeated condition.

One participant did state they believed that the arrays in the Interleaved condition were repeating but the much wider ranges in the Interleaved condition - compared to the Repeated - argues against this having been a common feeling.

## General Discussion

### *Heuristic Elicitation*

It seems reasonable to conclude that elicitation techniques enabling people to use the well-honed, heuristic judgment and decision tools already at their disposal are powerful tools for reducing bias. The degree of overconfidence observed in the MOLE responses was much smaller than in the traditional range elicitation conditions, particularly given that when asked for a wide confidence interval (80% plus), people tend to give ~50% intervals (Morgan & Henrion, 1990) and the MOLE generated a 100% interval.

Of greater interest is the fact that this method works not just by causing people to consider more values, thereby including a wider range of possibilities (i.e., increasing subjective accuracy by helping people realize the limits of their knowledge) but also by improving their objective accuracy. That is, while participants using traditional methods proved poor at estimating the true number of objects displayed, repeatedly judging which option was closer to the true value resulted in participants in the MOLE condition having a better idea of what that true value was.

### *Repeated vs Single Judgments*

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 - resulting in multiple searches of their beliefs about the stimulus. While it seems reasonable that this should increase the range of possibilities considered, it is unclear why this should improve accuracy, particularly given that the Interleaved condition, which should also result in multiple searches of a participant's beliefs, did not provide equal benefit.

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 better at making relative rather than absolute judgments. This supports the idea from the bounded rationality (Gigerenzer & Selten, 2001) literature that enabling people to answer questions in formats they are adept with is a good way to avoid bias in judgment and decision making.

Another interesting observation, arising from the results presented here relates to the differences in single judgments between Experiments 1 and 2. In Experiment 1, range estimates based on single judgments are very inaccurate – showing basically no correlation with the true value. In Experiment 2, by comparison, the mean correlations between the true value and the estimates made in each individual stage of the repeated judgment tasks are around 0.4. This difference, obviously, needs to be accounted for.

One possibility is that this is the result of noise reduction due to the increased time participants spend looking at each of the arrays in the second experiment – although the fact that the

effect exists of the Interleaved condition as well as the Repeated undermines this explanation. An alternate explanation lies in the difference between the two experiments in terms of the sample and procedure.

Specifically, the sample used for Experiment 2 was older and comprised of many more graduates and postgraduate students. As such, it is possible that this group were more intelligent and or more motivated than the undergraduate sample used in Experiment 1. The repetitive nature of Experiment 1, with 40 estimation tasks compared to just 3 in Experiment 2 may also have led to participants paying less attention to each individual task in the earlier experiment. Regardless of which explanation is preferred, however, the most interesting observation is that the MOLE technique prevents any adverse effect in its results – the MOLE results being consistent across the two experiments.

#### *Limitations*

There remain, however, several caveats regarding the current MOLE method. It could, for example, be argued that the MOLE gave participants an unfair advantage in that it limited the range of values that the contrast values could be selected from to between 0 and 400 (remembering that the true value was always between 100 and 300). Figure 4 shows that, in the non-heuristic conditions, a number of estimates lie beyond this range, meaning that participants had the opportunity to be more inaccurate than in the MOLE. That said, the vast majority (98.1%) of estimates from all other conditions fell within that range – it having been chosen as a reasonable estimate of the range of responses – so any effect from this would be limited. The other relevant result for this issue is the qualitative difference in the scatterplots, which shows better accuracy in the MOLE condition across the full range of stimuli.

Additionally, the need for bounds limits the usefulness of the MOLE, as currently formalized, to situations where limits can be put on what people might believe (although these initial limits could be set very broad without greatly increasing the number of steps required to complete the task due to the MOLE's iterative narrowing of the range to exclude infeasible values). There remains, however,

a risk of limiting the outcomes a person is allowed to choose. That said, this should not be a problem in many applied domains where expert knowledge is being sought - where there are, often, known limits on outcomes.

The time requirements for application of the MOLE also need to be considered. The procedure takes approximately 3 minutes per elicitation task – compared to a fraction of that for a traditional range elicitation. The improved accuracy and calibration, however, will more than outweigh this in most situations where expert opinion is being elicited as it is most commonly the case, in actual elicitations, that only a single value needs to be elicited from a given expert on a given occasion.

#### *Future Directions*

Given our findings, it seems worthwhile to continue looking at relative and repeated judgments as methods for avoiding bias in elicited responses. In addition to extending its use to non-visual elicitation tasks, an obvious direction is to refine the MOLE procedure such that it can automatically determine if a person has reached the limits of their certainty, rather than requiring a set number of questions.

Further development of the MOLE to address some of the concerns raised above and improve its ability to capture people's beliefs also 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.

The application of this approach to other biases that impact on elicited responses such as anchoring would also be of interest - given how resistant to debiasing anchoring has proved ((Wilson, Houston, Etling, & Brekke, 1996).

Finally, while we believe the focus on relative judgments is an important advance in developing elicitation methods, it should be possible to improve the way in which subjective PDFs are generated. The current method was chosen so as to minimize the assumptions needing to be made, but remains somewhat ad hoc in nature. One interesting possibility is to follow the recent lead of Sanborn and Griffiths (2008), who apply modern computational Bayesian sampling algorithms, based on Markov-Chain Monte Carlo methods, as experimental procedures for understanding the subjective probability distributions people use to represent mental categories. Applying the same principled ideas to the problem of value elicitation is a promising direction for future research.

### *Conclusion*

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 and providing estimates superior to the repeated judgment methods explored herein and greatly superior to traditional range estimates.

Heuristic-based elicitation methods like the MOLE thus seem to be a worthwhile addition to the arsenal of researchers interested in reducing the impact of bias on elicited responses. While the fine detail requires further refinement, the basic premise, of using relative rather than absolute judgments, is strongly supported by the findings herein and the concept seems well placed to contribute to our understanding of how our cognitive abilities give rise to bias and to aid in improving the accuracy of forecasting in a variety of areas.

## References

- Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49, 188-207.
- Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded Rationality: the adaptive toolbox* (Vol. 84). Cambridge: Massachusetts: MIT Press.
- Gigerenzer, G., & Todd, P. M. (Eds.). (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.
- Hawkins, J. T., Coopersmith, E. M., & Cunningham, P. C. (2002). *Improving stochastic evaluations using objective data analysis and expert interviewing techniques*. Paper presented at the Society of Petroleum Engineers 78th Annual Technical Conference and Exhibition, San Antonio, Texas.
- 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.
- Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85-126.
- Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 1038-1052.
- 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.
- Morgan, M. G., & Henrion, M. (1990). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.

- Morgan, M. G., & Keith, D. W. (1995). Subjective judgements by climate experts. *Environmental Science and Technology*, 29(10), 468A-476A.
- Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26(1), 41-47.
- Russo, E. J., & Schoemaker, P. J. H. (1992). Managing Overconfidence. *Sloan Management Review*, 33, 7-17.
- Sanborn, A. N., & Griffiths, T. L. (2008). Markov Chain Monte Carlo with People. In B. Scholkopf, J. Platt & T. Hoffman (Eds.), *Advances in Neural Information Processing* (Vol. 20). Cambridge: MA: MIT Press.
- 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*. New York, NY: Random House.
- 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(7), 645-647.
- Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2007). Efficacy of bias awareness in debiasing oil and gas judgments. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Meeting of the Cognitive Science Society* (pp. 1647-1652). Austin, Texas: Cognitive Science Society.
- Welsh, M. B., Begg, S. H., Bratvold, R. B., & Lee, M. D. (2004). SPE 90338: Problems with the elicitation of uncertainty. *Proceedings of the Society of Petroleum Engineers 80th Annual Technical Conference and Exhibition, Houston, Texas: SPE*.
- Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4), 387-402.

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: how to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(6), 1167-1175.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *The Journal of Economic Perspectives*, 18(2), 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.

Author Note

Matthew B. Welsh, Australian School of Petroleum, University of Adelaide, Adelaide, South Australia 5005, Australia; Michael D. Lee, Department of Cognitive Sciences, University of California, Irvine, CA 92697, USA; Steve H. Begg, Australian School of Petroleum, University of Adelaide, Adelaide, South Australia 5005, Australia.

MBW and SHB are supported by ExxonMobil and Santos through their support of the CIBP at the Australian School of Petroleum.

This paper is partially based on work previously presented at the Cognitive Science Conference.

The authors wish to thank Ben Schultz and Carolyn Chen for their assistance in data collection, Dan Navarro for his assistance with the code and Danny Oppenheimer and Mark Steyvers for their comments on earlier versions of this manuscript.

Correspondence concerning this article should be addressed to: Dr M.B. Welsh, Australian School of Petroleum, University of Adelaide, North Terrace, Adelaide 5005, South Australia, Australia. Email: [matthew.welsh@adelaide.edu.au](mailto:matthew.welsh@adelaide.edu.au)

## Footnotes

<sup>1</sup> Given previous research indicating that people are better at evaluating than generating confidence intervals, in all of the traditional range elicitation methods, people were, rather than being asked to generate confidence intervals of particular types (e.g., 90% confidence intervals), instead asked to give a minimum and maximum value and then indicate how confident they were that the true value would fall within that range. That is, allowance was made, e.g., for the fact that people would not regard their “minimum” as the true minimum.

Table 1

*Ordering of Elicitation Methods*

Group	Elicitation Methods			
A	1	2	3	4
B	4	3	2	1
C	2	4	1	3
D	3	1	4	2

Note: 1=Simple, 2=Triangular, 3=Iterative, 4=MOLE

Table 2

*Hits and mean confidence rating by condition*

Condition	Hits	Trials <sup>#</sup>	% Hits	Confidence
Simple	92	340	27.1%	75.5%
Triangular	81	340	23.8%	72.6%
Iterative	97	340	28.5%	72.7%
MOLE	308	340	90.6%	100% *

# - 20 of the 360 trials were excluded as individual analyses indicated that participants had either misinterpreted the experimental instruction or were deliberately answering incorrectly in order to limit their participation time. \* - assumed confidence level.

Table 3

*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 4

*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.

Table 5

*Time to complete task by condition*

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

Figure Captions

*Figure 1.* MOLE GUI

*Figure 2.* Subjective PDF construction process.

*Figure 3.* Mean width of range by elicitation method.

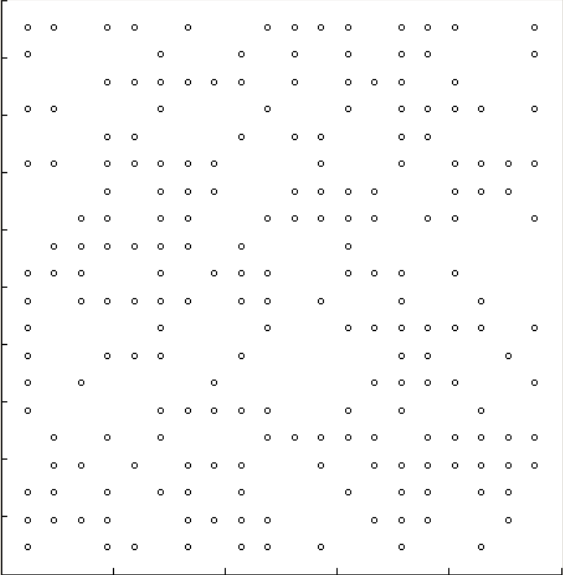
*Figure 4.* Scatterplots comparing number of objects with the estimated mode for each elicitation condition.

*Figure 5.* Revised MOLE GUI.

*Figure 6.* Example of PDF construction from individual, uniform PDFs.

*Figure 7.* Scatterplots of true and estimated number of circles in arrays.  $N = 36$  in all cases.

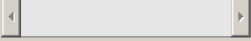
How many circles are displayed below? Please press the button below the number that is closer to your estimate.



177      195

Select 1      Select 2

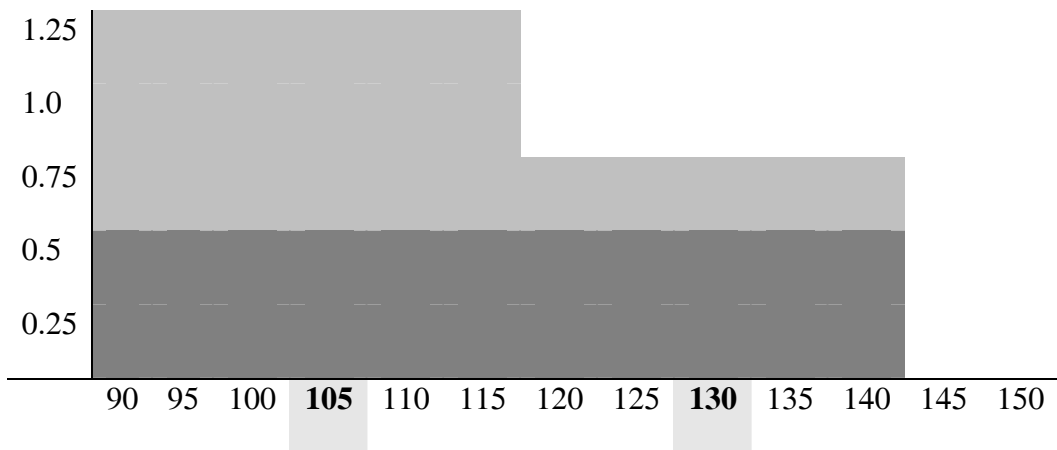
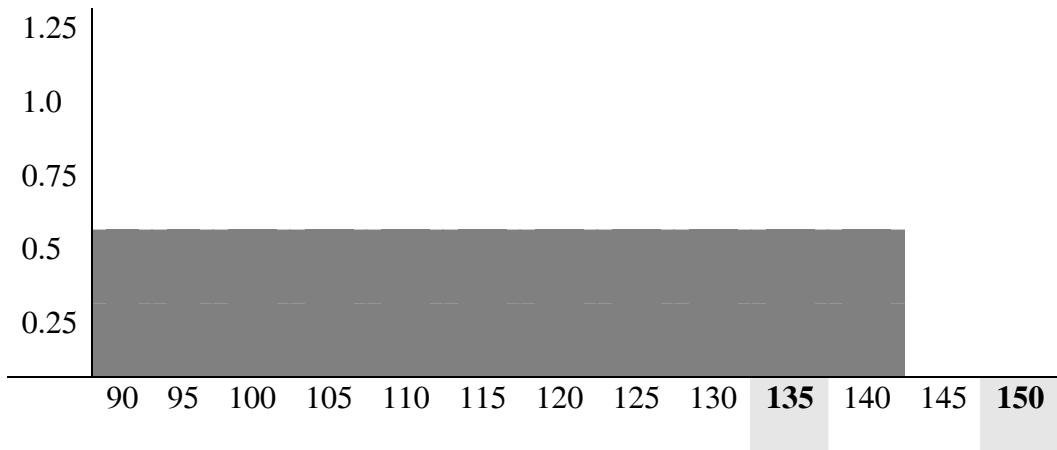
50%      100%

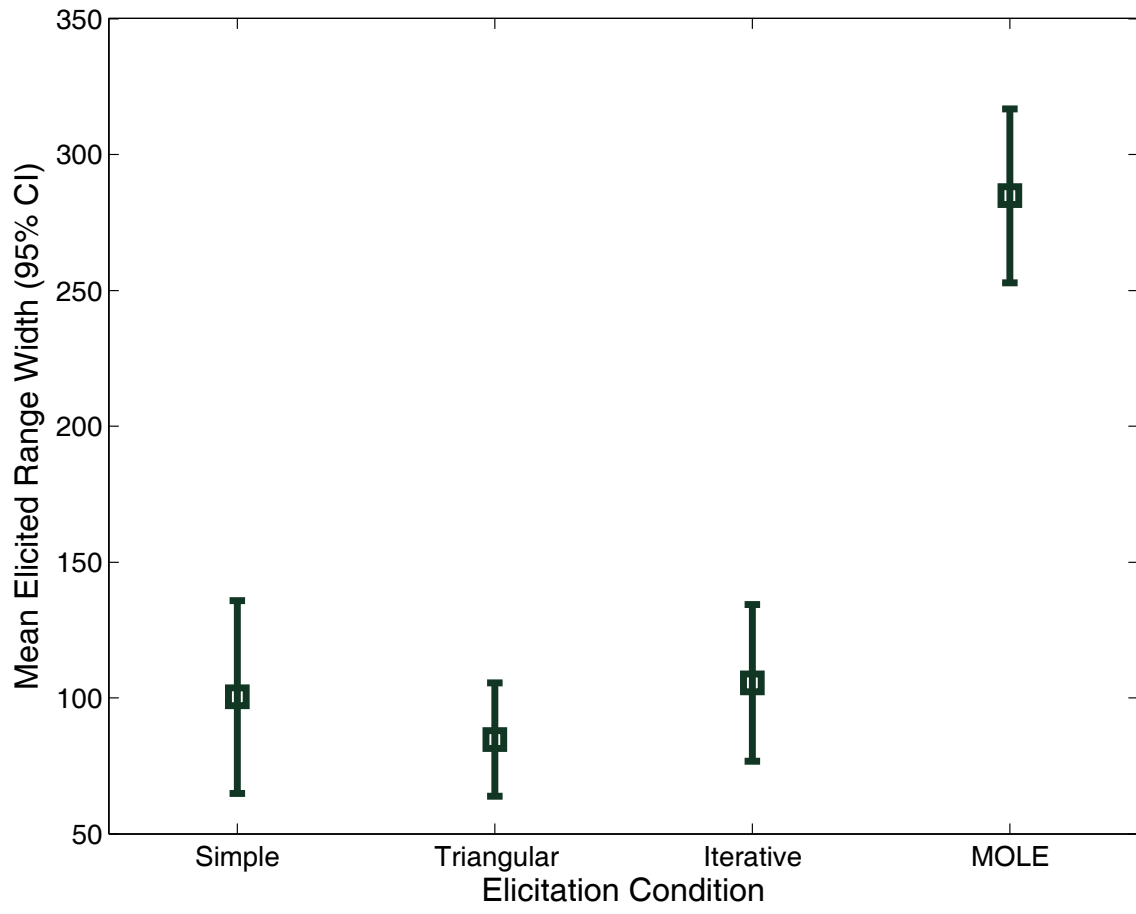


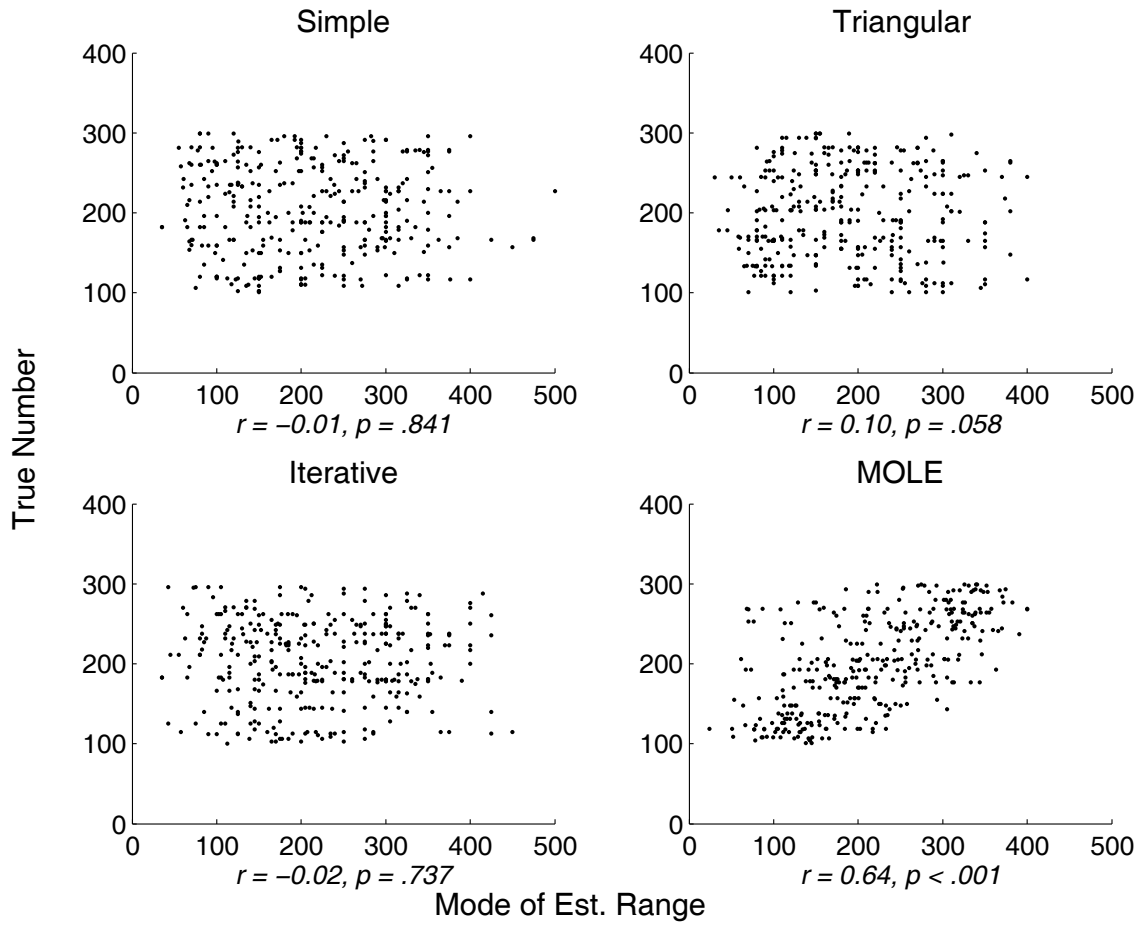
50

NEXT

NEW

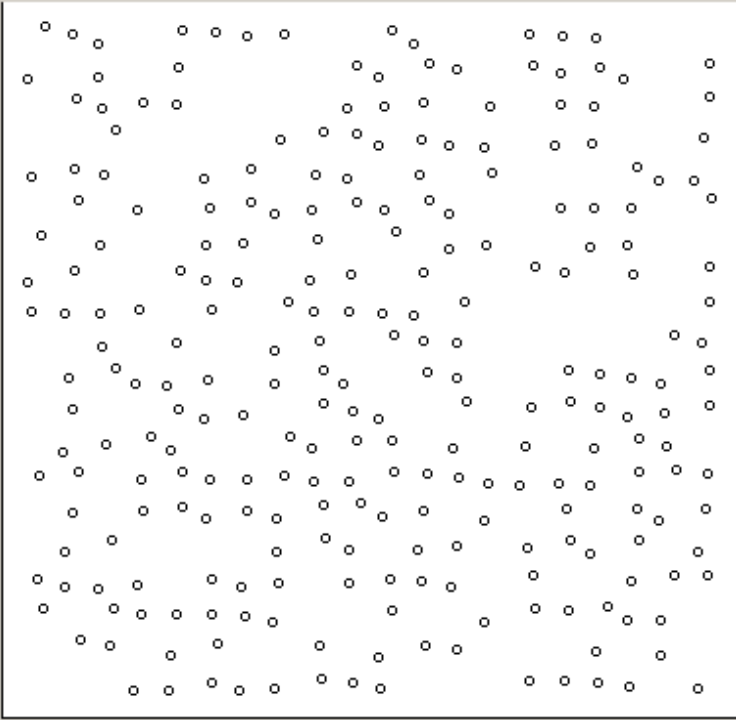






mole09

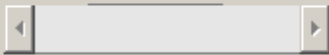
How many circles are displayed below? Please press the button below the number that is closer to your estimate.



78      115

Select 1      Select 2

50%      100%



50

NEXT

NEW

