More-Or-Less Elicitation (MOLE): reducing bias in range estimation and forecasting

Matthew B. Welsh and Steve H. Begg
Australian School of Petroleum
University of Adelaide
North Terrace, Adelaide, SA 5005
Australia

E: matthew.welsh@adelaide.edu.au
Ph: +61 411 249 303
Fax: +61 8313 8030

Acknowledgements: The authors thank past and present Centre for Improved Business Performance (CIBP) research club members – BG Group, ExxonMobil, Santos and Woodside – for supporting this research through the Australian School of Petroleum, University of Adelaide. We also thank Michael Lee for his input to previous versions of the MOLE and him and two anonymous reviewers for their comments on this manuscript. Some data described herein were presented at the Cognitive Science Conference and included in its (non-archival) proceedings (Welsh & Begg, 2015; Welsh, Lee, & Begg, 2008, 2009) but the description of the MOLE and analyses have been significantly updated and expanded herein - and an additional experiment incorporated.

Abstract

Biases like overconfidence and anchoring affect values elicited from people in predictable ways – due to people's inherent cognitive processes. The More-Or-Less Elicitation (MOLE) process takes insights from how biases affect people's decisions to design an elicitation process to mitigate or eliminate bias. MOLE relies on four, key insights: 1) uncertainty regarding the location of estimates means people can be unwilling to exclude values they would not specifically include; 2) repeated estimates can be averaged to produce a better, final estimate; 3) people are better at relative than absolute judgements; and, 4) consideration of multiple values prevents anchoring on a particular number. MOLE achieves these by having people repeatedly choose between options presented to them by the computerised tool rather than making estimates directly, and constructing a range logically consistent with (i.e., not ruled out by) the person's choices in the background. Herein, MOLE is compared, across four experiments, with eight elicitation processes – all requiring direct estimation of values – and is shown to greatly reduce overconfidence in estimated ranges and to generate best guesses that are more accurate than directly estimated equivalents. This is demonstrated across three domains – in perceptual and epistemic uncertainty and in a forecasting task.

Keywords: bias; elicitation; forecasting; overconfidence; range estimation; anchoring.

1. Introduction

Elicitation describes the conversion of experts' subjective beliefs into probabilities to be used in modelling and forecasting; in effect, extracting other people's knowledge to reduce our own uncertainty regarding a future event or unknown state of the world (Wolfson, 2001). As such, it is essential for industries dealing with high uncertainty such as pharmaceuticals and petroleum exploration and development – the latter of which has been described as a classic example of decision making under uncertainty given high up-front investments and low probabilities of economic success for new projects (Newendorp & Schuyler, 2000).

1.1 Problems for Elicitation

Unfortunately, decades of psychological research, including seminal work by Tversky and Kahneman (Tversky & Kahneman, 1973, 1974, 1981) have shown the values we elicit from experts can be biased as a result of the ways in which people typically think – that is, our cognitive limitations and processes.

Key amongst these are: overconfidence in range estimation (hereafter 'overconfidence'); and bias arising from the anchoring-and-adjustment heuristic ('anchoring'). The first describes the tendency for ranges that a person believes (to a stated level of probability) will contain a future or unknown value to be too narrow – with the result that these ranges contain the true value less often than the person's stated confidence would suggest (Lichtenstein, Fischhoff, & Phillips, 1982). The second describes people's tendency to base estimates on any number currently at hand, regardless of its relevance – including random numbers (Tversky & Kahneman, 1974). While these are far from the only biases, they are central to the field and feature on many lists of important biases affecting human decision making (see, e.g., Piatelli-Palmarini, 1994; Russo & Schoemaker, 2002; Thaler & Sunstein, 2008). Both also have clear implications for values elicited from experts.

1.2 Psychological Basis of Bias

To understand how to design elicitation processes that avoid or limit the impact of biases, it is necessary to understand how the biases arise – that is, which cognitive limitations or tendencies are implicated in their appearance.

1.2.1 Overconfidence

Overconfidence, as discussed herein, is Moore and Healy's (2008) 'overprecision' –

the tendency for elicited ranges to contain predicted values less often than people expect (Lichtenstein et al., 1982). For example, when people are asked to provide ranges that have an 80% chance of capturing the true value, typically these ranges capture the true value less than 50% of the time. This has been observed: in both expert and novice samples across various fields (see, e.g., Morgan & Henrion, 1990; Welsh & Begg, 2016); and in observations of actual oil industry predictions (Hawkins, Coopersmith, & Cunningham, 2002). That is, people's confidence judgements tend to be miscalibrated because their ranges are too narrow.

This has important implications for decision making as uncertainty in outcomes determines whether additional funds should be spent on uncertainty reduction or risk mitigation/upside capture strategies. As such, this bias can have multi-million dollar impacts on investment decisions (see, e.g., Welsh, Begg, & Bratvold, 2007).

While some authors have argued that overconfidence results from the differences between people's inherent, frequentist understanding of probability and the need, in elicitation tasks, to state subjective probabilities for unique, non-repeatable events (Gigerenzer, 1991; Gigerenzer, Hoffrage &Kleinbolting, 1991) this Brunswikian approach suggests that overconfidence is entirely artefactual and will disappear in evaluation tasks that allow a person to construct a reference class by asking a different question – "how often have ranges that I set actually contained the true value?" - that can be answered using natural, frequentist reasoning. In fact, while there is evidence that people are better at evaluating than generating ranges (Winman, Hansson, & Juslin, 2004), range evaluation still results in some overconfidence (Winman et al, 2004) and recent work by Ferretti, Guney, Montibeller and von Winterfeldt (2016), found limited benefit in an experiment where participants both generated and then evaluated a range.

Given this, overconfidence can not be dismissed. Instead, other causes and debiasing strategies need to be considered. Research on this, however, has shown that overconfidence is resistant to simple debiasing attempts – such as exhortations to increase the widths of ranges (Lichtenstein et al., 1982) or awareness of the bias' effects (Welsh, Begg, & Bratvold, 2006) and indicates that people are resistant to providing ranges as wide as would be necessary to capture their true uncertainty because such ranges are deemed uninformative (Yaniv & Foster, 1997).

This reflects a possible cause of overconfidence – the informativeness-accuracy trade-off (IAT; Yaniv & Foster, 1995) - whereby people are argued to deliberately provide narrow ranges because their preference is to be informative over accurate (well-calibrated). This

explanation bears some resemblance to one raised in section 1.2.2, below, regarding the mode of operation of the anchoring bias wherein people stop adjusting their estimate once they reach their region of uncertainty. That is, continuing to adjust one's estimate of a range's endpoint further into the region of uncertainty decreases the informativeness of the range and does not increase the likelihood of the end-point itself being true.

This possible connection between anchoring and overconfidence echoes the original hypothesis that anchoring caused overconfidence through people anchoring on their best estimate and failing to adjust far enough away from it (Tversky & Kahneman, 1974). This does seem to occur in some cases (Heywood-Smith, Welsh, & Begg, 2008; Russo & Schoemaker, 1992) but not in others (Block & Harper, 1991; Bruza, Welsh, Navarro, & Begg, 2011; Welsh, Begg, Bratvold, & Lee, 2004), suggesting the relationship is complex but that approaches designed to avoid bias in this way could sometimes be beneficial.

A demonstrated way to alter overconfidence is by changing elicitation format. For instance, there is evidence that splitting the task into parts – asking for the 10^{th} and 90^{th} percentiles separately rather than for a range that a person is 80% confident will contain the true value – produces better results, possibly as a result of lifting limitations on cognitive effort (by splitting a single task into two tasks) (see, e.g., Juslin, Wennerholm, & Olsson, 1999). Other debiasing techniques, proposed in Montbellier and von Winterfeldt's (2015) and tested in Ferretti et al. (2016) include the use of bets to identify errors in probability and the presentation of counterfactuals (values lying outside the initial range). Both were found to have only small effects on the width of elicited ranges although the use of bets improved the best estimate. That is, despite debiasing techniques that offer some benefit, overconfidence remains a significant problem for elicited values.

1.2.2 Anchoring

The original description of anchoring-and-adjustment argues it results because people use the anchor as a starting point for their estimation process and then adjust away from the anchor until they reach a point at which they feel no need to adjust further (Tversky & Kahneman, 1974). This, it is argued, leads to bias because the point at which a person will stop adjusting is the point nearest to the anchoring value out of their range of feasible values.

This explanation is supported by research but so is a second explanation based around the idea of *priming* (for a recent discussion of the two explanations, see Furnham & Boo, 2011). This holds that the anchoring value sets the region of possible values a person will

start evaluating. For example, if asked whether Mt Everest is higher than 5000 metres, a person might start by considering whether 5000 metres is a reasonable estimate and, only if they decide it is not, will they start to draw possible cues (e.g., the heights of other mountains starting with those nearest to 5000m) from their memory for consideration.

Regardless of the cause, anchoring poses a problem for elicitation in that: any prior number can affect a person's estimate (Tversky & Kahneman, 1974); the effect is robust (Mussweiler, 2002; Mussweiler, Strack, & Pfeiffer, 2000); it affects novices and experts (Northcraft & Neale, 1987); and is not reduced by people's awareness of the effect (Welsh et al., 2006). This implies that any elicited value could be affected by previously seen or experienced values – no matter how irrelevant these might be.

1.3 Building Better Elicitation

The central reason for understanding the mode of action of biases is, of course, to assist in avoiding those biases. That is, understanding what gives rise to a particular bias allows us to avoid the bias by avoiding those circumstances. The following sections describe how biases and other quirks of human cognition can be used to improve elicitation.

1.3.1 Retaining Uncertainty

The observation (from Kahneman, 2011) noted in section 1.2.2, that one cause of anchoring bias is that people stop adjusting once they have reached a possible estimate that lies within their region of uncertainty, has important implications for how a range should be elicited from an individual. Specifically, it implies that if a person is allowed to construct a range by starting at their best estimate and working out towards the ends from there, they will tend to stop at the *inner* edge of their regions of uncertainty for both the high and low points and, as a result, produce a range that is narrower than they otherwise might – as suggested in the description of overconfidence in section 1.2.1 and illustrated in Figure 1.

Given this, the obvious solution is to have people construct their ranges in the converse fashion – starting with a very (i.e., much *too*) wide range and asking them to adjust the endpoints inwards, removing regions they are certain the true value will not fall within. In this way, a person's uncertainty could be used to preserve the range as they would, presumably, stop cutting away portions of the overall range as soon as they reach the outside edge of their region of uncertainty.

1.3.2 Relative Judgements

A robust finding from psycho-physics is that people perform better when asked to make relative rather than absolute judgements. For example, Stroop (1932) showed people could very accurately sort weights into order but were poor at estimating absolute weights. S imilar effects have been shown in many perceptual tasks (see, e.g., Miller, 1956).

Unfortunately, of course, most elicitation are undertaken in order to obtain absolute values to use as forecasts or estimates of unknown parameters. There is, however, evidence that people use relative judgements to construct their absolute estimates (Stewart, Brown, & Chater, 2005) – a finding that echoes the observations about the priming explanation for anchoring in section 1.2.2 where people are argued to draw possible values from memory for comparison with the anchor. The difference in accuracy between relative and absolute judgements therefore suggests that additional bias results from this translation process.

In light of this, it seems valuable to consider elicitation processes that allow people to make relative judgements rather than absolute ones – as such judgements are more likely to be correct and can then be translated into absolute judgements by an algorithm that produces less bias than human cognition, increasing the accuracy of the person's estimate.

1.3.3 Repeated Judgements

Another robust finding is the so-called 'wisdom of crowds' (Galton, 1907; Surowiecki, 2004) - the tendency for aggregated estimates from a group of people to be superior to the estimates of individuals within that group. This is, primarily, a mathematical effect – an observation that any non-systematic biases will differ in direction and magnitude between individuals and thus tend to average out. The psychology, of course, comes in when considering the extent to which people's biases are, in fact, non-systematic. For example, using the wisdom of crowds approach on a group who had all seen the same anchoring value would result in an average estimate biased towards that anchor. Given this, diversity of opinion and background and independence of information is the ideal situation for wisdom of crowds effects and the larger the group, the better the results tend to be.

For elicitation tasks, however, the pool of people able to meaningfully interpret a question is limited by expertise and confidentiality, with the result that wisdom of crowds is of limited use. There is, however, still a benefit to be gained from repeated judgements – even from a single individual.

Research (see, e.g., Herzog & Hertwig, 2009; Vul & Pashler, 2008) has shown that

asking an individual the same question repeatedly and averaging their responses produces a better estimate than simply taking their first estimate. The reasoning is that, where individuals do not remember the exact answer they gave previously, they will construct a new estimate each time they are asked. Given the limitations of human memory and biases resulting from specific elicitation circumstances, however, the set of information a person draws on will tend to differ on each occasion (see, e.g., Juslin, Winman, & Hansson, 2007) – meaning estimates will not be identical. To the extent that errors and biases differ non-systematically they will tend to cancel out (Surowiecki, 2004) and average estimates will be superior to individual ones. Of course, the research noted above demonstrated that longer periods of time between elicitations increased the independence of estimates and, thus, the benefit gained from this – a concern given that values being elicited from experts are often time sensitive.

In practical terms, then, the ideal elicitation process is one allowing a person's knowledge to be probed in such a way as to allow them to make estimates one after the other – with no time delay – while preventing them repeating back previous estimates. The final, combined estimate from such a process should be superior to that from any single-estimate process.

1.3.4 Avoiding Anchors

Anchoring bias has proven resistant to debiasing methods based on awareness of the effect (Welsh et al., 2006) – although some success has been observed with more directed debiasing attempts that lead people to consider values other than the initial anchor (Mussweiler et al., 2000). This, of course, requires that a person recognise (have pointed out to them) the anchoring value – which poses no problem in experimental tasks but is more difficult in real-world circumstances, where an anchor could be a random number the elicitee has just encountered or a subconscious intuition based on previous situations. While this may seem to be an unfair discounting of expertise (i.e., the expert's intuition), it should be noted that the oil industry (and other areas where elicitation of uncertainty is most commonly used) fail to meet the criteria established by Kahneman and Klein (2009) for when expert intuition can be relied upon to be accurate. That is, the environment does not have the regularity of decisions and feedback required for expert intuition to reliably develop.

Thus, to avoid anchoring, the best path seems to be extending Mussweiler et al.'s (2000) approach and using other values to, in effect, 'wash out' the impact of any, one, anchoring value. That is, the elicitation process needs to require that the elicitee explicitly

consider a number of values across the range of possibility – a strategy already incorporated into advice for avoiding bias by proponents of debiasing (Kahneman, Lovallo, & Sibony, 2011; Russo & Schoemaker, 2002). While any one value might, on its own, act as an anchor, the need to consider all of them is argued to prevent biased sampling from a particular portion of the range of possibility – as suggested by the priming explanation of anchoring – and also the process of simple adjustment from an anchor towards a person's region of uncertainty.

1.4 The MOLE: More-Or-Less Elicitation

The MOLE process, used in all of the experiments described hereafter, uses the four insights above to create a computerised tool that guides a person through an elicitation process designed to limit the impact of overconfidence and anchoring while, simultaneously, attempting to increase estimates accuracy.

The first step of the MOLE is the selection of its starting range – ideally by someone other than the user. As the MOLE relies on cutting away areas of the range not considered feasible, starting with a very wide range (wide enough to contain any reasonable estimate) is advised. Where natural limits exist (e.g., percentages or proportions), these are appropriate starting ranges. Otherwise, databases of prior outcomes can be used to inform the starting range – which should include all values previously seen plus a margin of error at either end to account for previously unseen low or high values.

Starting with this initial (wide) range, MOLE randomly draws pairs of values. These values are presented to the participant, who is asked which value they believe is *closer* to the true value (of whatever parameter is being elicited). Once a participant has selected one of the options, they are asked to indicate their confidence in this judgement on a 50% (guessing) to 100% (certain) scale using a slider.

The MOLE then uses the selection and confidence judgment to update the range from which future values will be drawn using a simple, logical rule – that is, if a person is 100% confident that their selected option is closer to the true value than the alternative is, then values closer to that alternative have, logically, been ruled such out. For example, if a person were shown the values 100 and 200 and selected 100 with 100% confidence, then the MOLE would no longer include any values above 150 (the midpoint of the two options) when drawing future options as the participant has 'stated' that the true value is definitely closer to 100 than 200.

The MOLE then draws two new values from the (possibly truncated) range and

repeats the process for a set number of iterations (10 in the cases described herein). In this way, the wide starting range is cut down to a narrower range containing only values that a participant has not, specifically, ruled out.

At the end of its 10 iterations, the MOLE uses the non-100% confidence ratings to generate a person's best guess, using the following assumption: a person's best guess from any single judgement lies between the two options presented to them. Specifically, that it lies *confidence*/100% of the way from the unselected option to the selected option. For example, having been shown the values of 100 and 200, a participant selects 200 with 70% confidence – that is, they are 70% sure that the true value lies closer to 200 than to 100. We operationalize this as indicating that their best guess lies 70% of the way from 100 to 200 – that is, at 170. Had they selected 100 with the same confidence, their best guess would be 130 (70% of the way from 200 to 100).

One such best guess is generated for each non-100% confidence judgement, excepting where the options being compared lie outside the final, *feasible* range described above – reflecting early trials where the participant had not yet cut away those portions of the starting range. In this case, the judgement is discarded as misleading. The remaining best guesses are then averaged to produce the participant's overall best guess.

The MOLE process thus enables us to use all four of the techniques described in section 1.3 for building a better elicitation tool. It is designed to limit bias by: 1) reducing overconfidence by requiring the participant to rule out rather than rule in regions of possibility; 2) collecting repeated measurements of a person's best estimate in such a way as to prevent a person from simply repeating their preferred answer; increasing accuracy by 3) allowing people to make relative rather than absolute judgements (i.e, not requiring them to directly make an estimate of the parameter being elicited); and 4) avoiding the impact of any single anchor or priming effect by requiring the participant consider 10 pairs of values selected from across a wider range of possible values than a person might otherwise generate.

1.5 Aims and Structure of this Paper

This paper presents a series of four experiments comparing the MOLE to various elicitation methods under a variety of conditions to determine what benefit it may provide. As such, our overall objective is to compare the accuracy and calibration achieved by the MOLE with similar measures obtained using alternative elicitation processes and determine whether it works in a variety of distinct, elicitation tasks.

Given the MOLE's design, underpinned by psychological theory, we predict it will produce better range estimates – in terms of calibration and the accuracy of best estimates from those ranges – than elicitation processes wherein people directly estimate values. That is, the primary focus of the experiments is on assessing the impact of the MOLE on reducing overconfidence – which is a function of range width and the accuracy of their placement.

Experiment 1a compares MOLE to three elicitation methods requiring participants to directly estimate the number of circles on a visual display (perceptual uncertainty). The direct elicitation methods include: 'Simple' estimation of high and low values to produce a range; 'Triangular' estimation of high and low values and a best guess; and 'Iterative', where a participant's initial, high-low, interval estimate is challenged with a value lying outside the range and the participant is then asked if they want to revise their range. This study also includes the paper's only direct analysis of the effect of anchoring within the MOLE task – examining whether the MOLE's initial options affect the final, best estimate.

Experiment 2a uses the same, visual estimation task and compares the MOLE with two alternative elicitation methods reliant on repeated measurement of the same individual's opinions – in an effort to determine whether the MOLE's performance results from repeated measurement or longer exposure to stimuli. These are: 'Repeated', where the participant is asked to repeatedly estimate the minimum and maximum number of circles ten times while looking at the same stimulus – in order to test whether simple length of exposure leads to better estimation; and 'Interleaved', where the same stimulus is presented ten times to participants for estimation but these trials are interleaved between 30 distractor trials – increasing independence between the elicited ranges.

Experiment 2 tests the effect of varying the MOLE's starting range on its performance relative to two other elicitation methods: 'MMM', or minimum, maximum and most likely; and 'Dialectical', which asks participants to give their range and then to consider that the true value lies outside that range and decide whether it is more likely to lie above or below their range before revising their minimum and maximum estimates. In all cases, the stimuli in this experiment were numerical questions of fact (i.e., epistemic uncertainty regarding, e.g., geography). Five questions have percentage answers and thus naturally bounded sets of options (0 to 100%) while two other sets of 5 questions are naturally unbounded. For these, the MOLE's starting range is set to either 0-200% or 0-500% (of the true value) to test how the starting range affects the MOLE's performance relative to the direct elicitation methods.

Experiment 3 extends comparisons between the MOLE and direct elicitation methods

to a more realistic task wherein future values of stocks, commodities and meteorological events were forecast 7 and 28 days into the future. The elicitation method used for comparison with the MOLE here is a direct estimation of minimum and maximum values.

Following the experiments, a general discussion overviews the findings, discusses the practical use of the MOLE, caveats and future research, before drawing overall conclusions.

2. Experiment 1a: Overconfidence in Perceptual Uncertainty 1

2.1 Aims and Objectives

The first experiment was designed to compare estimates elicited using the MOLE method with those achieved using direct range estimation methods – in terms of both the accuracy of best guesses and calibration of responses.

Three elicitation methods were chosen for comparison with the MOLE: a simple range estimation task, where participants gave minimum and maximum estimates; a 'triangular' estimation task, which required participants provide a best guess prior to estimating the range; and a two-stage 'iterative' elicitation task, to assess the impact of calling participants' attention to regions outside their initially estimated range. Both variants of the simple range elicitation method were selected in light of the evidence presented in section 1.2.2 that these might impact on the level of overconfidence observed.

The uncertain parameter being elicited in this experiment was the number of circles in a visual display. This was selected as the perceptual paradigm made it simple to conduct a within-subjects design. That is, because the task remained the same across trials and largely unaffected by knowledge, it allowed tasks of equal difficulty in each condition – whereas a more typical, almanac-style, epistemic uncertainty task requires matching of question difficulty for a within-subjects design to be feasible. While this limits external validity, the estimation process was noted by an oil industry professional to share characteristics of a petro-physical analysis method known as point-counting used to estimate the proportion of different elements of a rock type (M. Sykes, personal communication, 2007) and the later experiments (2 and 3, herein) extend the MOLE to more typical elicitation tasks.

An additional goal was a test of an assumption underlying the MOLE – that being that provision of a large number of values during elicitation would limit the impact of anchoring. Within the current design, this requires testing whether the first values provided by the MOLE act as anchors on participants responses.

2.2 Method

2.2.1 Participants

Participants were 40 undergraduate students from the University of Adelaide. Nine were excluded due to computer errors during testing or after examination of their responses revealed nonsensical responses, leaving 31 (9 male and 22 female) with a mean age of 20.1 (*SD* = 1.9). Participants received a $10 book voucher for their participation.

2.2.2 Materials

Four graphical user interfaces (GUIs) were developed to enable automated testing of participants using each elicitation method. All 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 elicitation technique, the same basic GUI layout was used, with only the questions being asked and the response buttons differing. For example, Figure 2 shows the layout seen during More-Or-Less Elicitation (MOLE) condition, asking participants to select which of two values is closer to their estimate. GUI controls were sequentially locked and unlocked to ensure that participants answered each question before continuing to the next.

2.2.3 Procedure

Over the course of an hour, participants completed 10 trials under each of the four elicitation conditions - after being sorted at random into four groups to allow counterbalancing for possible order/learning effects, as shown in Table 1.

Simple Elicitation. Here, 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% (NB – while a min to max range should reflect a 100% confidence interval, this was included as a check of whether participants genuinely considered the range they generated to be such). A person's best guess in this task was estimated as the mid-point of their elicited range.

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.

Table 1. Ordering of Elicitation Methods

Group	Elicitation Methods			
A	S	T	I	M
B	M	I	T	S
C	T	M	S	I
D	I	S	M	T

Note: S=Simple, T=Triangular, I=Iterative, M=MOLE. These four orders were selected from the 24 unique possibilities as they form a balanced Latin Square, ensuring that every elicitation method is preceded and succeeded by every other once only.

Iterative Elicitation. Here, 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 - described as having been elicited from "previous participants" but actually calculated by the program to lie outside their own range (60% of their initial minimum and 140% of the maximum). Participants were then given the chance to adjust their minimum and maximum estimates. Once happy with their estimates, they were asked to indicate their level of confidence that the true value would fall inside their final range on a 0 to 100% range. As with the simple method, a person's best guess in this task was taken as the mid-point of their (final) elicited range.

More-Or-Less Elicitation. In the MOLE condition, participants did not directly estimate values. Rather, as described in section 1.4, 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, with the respondent's confidence ratings used to determine the range of feasible values (i.e., those the participant's answers did not rule out) and a person's best guess.

2.3 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 in terms of both overall calibration/overconfidence and the accuracy of their estimates.

2.3.1 Overconfidence

Overconfidence is defined here as the difference between the expected and observed proportion of occasions when the range contains the true value. Figure 3 shows the average score out of 10 achieved by participants in each of the four conditions.

It is clear from Figure 3 that all three techniques requiring participants to directly estimate ranges resulted in less than 30% of ranges containing the true value – with a comparison of the 95% confidence intervals around the means indicating little difference between them. By comparison, the MOLE, resulted in ~85% of ranges containing the true value.

In all cases, the assumed confidence for comparison with these hit rates is 100% - as participants were asked for minimum to maximum ranges – yielding overconfidence scores of between 72.9 and 77.7% for the three standard elicitation processes and 15.2% for the MOLE and means that, for analyses the overconfidence and hit rates can be used equivalently.

While the magnitude of the differences in Figure 3 renders it moot, a repeated measures ANOVA was conducted, confirming 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 were conducted, post-hoc, for each unique pair of elicitation methods to determine which conditions were driving the significant ANOVA result. These indicated that only the MOLE condition differed significantly from the others, $t(30) = 19.2$, 16.9 and 14.9 (from the Simple, Triangular and Iterative, respectively), $p << .001$ in all cases.

However, as noted above, there is some doubt that people interpret minimum and maximum labels as strongly as this when generating ranges and, as such, comparisons with people's evaluations of their own ranges were also made for the three standard elicitation processes. People's confidence that their range would contain the true value was: 73.8% (Iterative); 73.9% (Triangular); and 76% (Simple). Combining these values with the observed hit rates yields overconfidence scores of between 48.9 and 51.6%. (NB – the confidence level in the MOLE condition is assumed to be 100% as this method did not include direct rating of

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.) Using these overconfidence scores instead of those described above did not change the overall results, however – with the MOLE still producing significantly better calibration and the other three methods being largely equivalent.

This appears to be driven largely by the difference in range widths between the conditions with the Simple, Triangular and Iterative methods all producing similar range widths (77.9, 71.9 and 83.6, with *SD* = 54.3, 45.0 and 59.7, respectively), while the MOLE produced significantly wider ones (263.9, *SD* = 132.7). That is, people in the direct elicitation methods produced ranges that were far too precise (given what they actually knew). In the MOLE, by contrast, while people reduced the width of their range from the starting point (i.e., 400), they tended not to do so by a large amount.

Accuracy. To assess the objective accuracy of participants' responses under each elicitation condition, the best guess from each elicited range (as described above) was compared with the true value. Scatterplots showing these data are included as Figure 4.

Figure 4 suggests that only in the MOLE condition did participant estimates accurately track the number of objects in the stimuli. Across the 310 datapoints (31 participants by 10 trials), the correlation between the means of the estimated range and true values was moderately high and highly significant, $r = 0.64$, $p < .001$, whereas correlations between the true values and the remaining elicited means were all near zero, r = -0.10, .06 and 0.01 for the Simple, Triangular and Iterative method respectively, $p > .05$ in all cases. This analysis, however, was performed across all data points, meaning that differences in individual skill might reduce or obscure any correlation. Thus, Figure 5 shows the distribution of individual participants' correlations. Analysis at this, individual, level however, yielded similar results, with the participants' median correlation in the four conditions ranging from 0.79 (IQR=[.53 .90]) for the MOLE to .07 (IQR=[-.13.24]) for Triangular, .01(IQR=[-.31 .12]) for Simple and -.02 (IQR=[-.34 .21]) for Iterative.)

2.3.2 Anchoring in the MOLE

While no specific anchors were included, the first values displayed by the MOLE have the potential to act as such. If this occurred, one would expect a positive correlation between one or both of the initial values and the best estimate generated by the MOLE. To test this, correlations were calculated between each of the first pair of values displayed (the

low and high option) and the best estimate generated by the MOLE process for each of the 40 participants across the 10 elicitations undertaken using the MOLE.

In both cases, the median correlation between the best guess and first value was close to zero, r = .00 and .09, *IQR* = [-.30 .48] and [-.13 .33] for the initial low and high values respectively. Binomial tests confirmed that the number of positive correlations did not differ from what would be expected by chance alone: 20/40 for the initial low value, p = .563 (one-tailed); and 24/40 for the initial high value, p = .134 (one-tailed).

2.3.3 Other Findings

Best Guesses and Overconfidence. As noted above, the triangular method was included to determine whether requiring participants to give a best guess prior to fixing their confidence interval's end-points would affect its width and thus their levels of overconfidence – as previous research on this question has been mixed.

Looking at the data in Figure 3, however, one sees little difference between the hit rates provided in the two conditions of interest (Simple and Triangular). Participants in the Triangular condition did give, on average, narrower ranges (M = 84.7, SD = 61.8) than they did in the Simple condition (M = 100.3, SD = 105.0) but the analyses above indicate no significant difference between overall performance in terms of overconfidence.

Iterative Elicitation. The Iterative method was included to see 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 Figure 3 however, one sees little evidence in line with expectations - participants' performance in two conditions being near identical.

2.4 Conclusions

Our results show a clear benefit of the MOLE technique for both the calibration and accuracy of elicited ranges. We found little support, however, for the role of initial best guesses or simple counter-intuitive values in improving elicitations – the latter observation being in line with Ferretti et al.'s (2016) results.

Further – and in line with the assumption underlying the MOLE's design - the values provided at the beginning of the task have no discernible effect on the final estimates. That is, there is no evidence of participants anchoring on either of the first pair of values seen. While not a direct test of the MOLE's proposed mechanism for avoiding anchoring bias (i.e., the

provision of multiple values), the absence of the typically robust (see, e.g., Mussweiler, 2002) anchoring effect in our results would seem to lend the hypothesis support.

2.4.1 Caveats

There are, however, some caveats regarding the MOLE method as used in Experiment 1a. Firstly, the MOLE process necessarily resulted in participants spending 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 overconfidence. The MOLE also requires more effort per trial, which resulted in more participants being excluded based on their failure to sensibly complete the MOLE than the other conditions. (Of course, this is, unlikely to cause a problem in applied settings where experts are undertaking tasks relevant to their roles and where multiple parameters tend not to be elicited simultaneously as was done for the purposes of the experiment).

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

A third concern relates to the best estimates calculated in the Simple and Iterative conditions from the mid-point of the participant's range - assuming a symmetrical distribution. In fact, the Triangular data showed some right skew with 53% of ranges extending further towards the high side, 19% symmetrical and 38% extending further to the low side. This suggests that best estimates for the Simple and Iterative conditions might be better modelled assuming a non-symmetrical distribution. In practical terms, however, this seems relatively minor as the Triangular data was not significantly more accurate than these alternatives.

Finally, the very poor performance of participants on the non-MOLE tasks warrants comment as the lack of correlations between estimate and actual values suggests that they either found the task extremely difficult or were unmotivated. (It should be noted, though, that the observation that the MOLE produced viable estimates even under such trying circumstances supports the idea of an elicitation process based around how people are best able to make judgements.)

At an individual level, some participants did show some evidence of better estimation with correlations ranging up to 0.82 in the Iterative condition but negative correlations

seemed as likely overall. Beyond questions of ability and motivation, it is also possible that this could results from people revising their order of magnitude part-way through the task– for example, if a person, on beginning the task, thought that estimates in the 100-200 range were appropriate but then, after seeing several trials, changed this to estimates in the 300-400 range, this could result in a set of responses with a high and a low cluster of estimates – each having a positive correlation within it but showing no overall correlation because the high estimates in the low cluster are lower than the low estimates of the high cluster.

This could account for the few outlying values observed in Figure 4 where some estimates above 400 were observed in the direct elicitation methods; which were prevented by the MOLE's preset range of 0-400. These values, while rare and having no overall effect on the accuracy of estimates in the direct elicitation methods, could reflect instances where people changed the magnitude of their responses.

3. Experiment 1b: Overconfidence in Perceptual Uncertainty 2

3.1 Aims and Objectives

The results of Experiment 1a supported the idea that the MOLE is superior method to traditional range estimation. There were, however, questions arising out of the results – 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 typical problems with repeated judgements from an individual. 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 repetition or the combination of the MOLE's four underlying principles (repetition, relative judgments, multiple values to foil anchoring and the 'outside-in' range construction) that provides the benefit. It is also necessary to assess whether the benefit of 'repeated measures' in the MOLE results simply from the additional time spent by participants examining a stimulus figure laid out in neat rows and columns.

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 - 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 solely from noise reduction due to increased time spent on each elicitation task.) It was hypothesised that the MOLE would still provide a benefit over and above that yielded by repetition with distraction which would, in turn prove superior to simple repetition.

A secondary adjustment to the design from Experiment 1a was to include a small amount of noise in the location of the circles in the stimulus display, so as to prevent their forming lines.

3.2 Method

3.2.1 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 received a $10 book voucher for their participation.

3.2.2 Materials

As in Experiment 1a, graphical user interfaces (GUIs) were designed - one for each experimental condition - displaying a random array of between 100 and 300 circles at each trial but differing in terms of the responses available to participants.

Figure 6 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. The other two GUIs, “Repeated” and “Interleaved”, were variants on the Simple method described in Experiment 1a. The primary difference between these and the MOLE GUI was that, rather than selecting presented alternatives, participants were asked to enter minimum and maximum estimates for the number of circles into editable text boxes. They then rated how confident they were that the true value would fall in that range using a 0-100% slider.

3.2.3 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 design. Participants were allowed a short (2 minute) break between conditions while the experimenter checked that the data had saved and started the next part of the experiment. A single trial was conducted under each

condition and most participants completed the task in less than 40 minutes; none taking more than an hour.

Mole Procedure. The MOLE GUI worked exactly as described in Experiment 1a with the exception that participants here completed only a single trial.

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. After this, participants were also asked to give a confidence rating for how likely it was that the true value would fall within the range they had just generated.

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 stimulus displays rather than just one. Specifically, forty arrays of between 100 and 300 circles were presented and participants were 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 repeat arrays were distributed in a pseudo-random manner throughout 30 distractor trials to prevent participants seeing two identical arrays immediately adjacent or noticing any simple pattern (i.e., not every fourth trial). By interleaving the experimental trials amongst distractor trials, it was expected that some problems with repeated judgment could be overcome.

3.3 Results

3.3.1 Data Manipulation

Outlier Removal. During analysis, discrepancies were observed between a participant's statements regarding their beliefs (made during testing) and the estimates recorded by the GUIs. Specifically, 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 accidentally entered the wrong value. To prevent this and other, unnoticed, errors from impacting results, all participants' data were analysed 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.

Estimated Range: For the MOLE, a person's estimated range was calculated as described for Experiment 1a. For the other conditions, their final range was taken to run from the lowest minimum value they provided on any trial to their highest maximum.

Best Estimates. Participants' responses were used to generate their best estimates as well as their intervals. The process used to generate the best estimate from the MOLE data was exactly as described above.

In the Repeated and Interleaved conditions, by comparison, a somewhat simpler (although related) method of best guess calculation was used. As each participant had estimated 10 ranges (Minimum to Maximum) for a given stimulus, the participant's overall, best guess was taken to be simply the average of the midpoints of their ten ranges.

3.3.2 Comparison of Elicitation Methods

To compare elicitation methods a number of measures are required - assessing both 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 assumed confidence level of 100%. (NB – as was the case in Experiment 1a, participants evaluated the chances of their own ranges in the Repeated and Interleaved conditions containing the true value – at 74.3% and 73.1% on average – but these ratings apply to individual ranges rather than the final, composite range.) Table 2 summarizes these key statistics across elicitation techniques.

Table 2. Summary of elicitation technique performance.

	Accuracy		Calibration	
Technique	r	|% Error|	% Hits	Confidence
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

Accuracy of Elicitation Methods. Figure 7 shows scatterplots between estimates made in each condition and the true value. Looking at this, one can see that estimates from all conditions show evidence of some degree of accuracy – with a positive correlation between the estimates and the true value, varying 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 $p < .001$, suggesting that estimates elicited using the MOLE may be better predictors of the true value (although, given the small sample, these correlation coefficients are not statistically distinguishable)

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 column 2 of Table 2, one sees the percentage error scores for participants in each elicitation method. 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 ANOVA, conducted comparing these results, found a significant result, $F(2,70) = 2.41$, $p = .016$ (one-tailed). Paired sample t-tests were used, post-hoc, to identify the conditions driving this results. These indicated that, the MOLE and Interleaved methods produced better results than the Repeated, $t(35) = 1.81$ and 1.70, $p = .020$ and .025 (one-tailed), respectively.

Calibration. In all three conditions, participants made confidence judgments after every individual judgement (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 responses. 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 2.

Looking at Table 2, one sees the MOLE produced the best calibrated results, with 91.2% of the composite ranges containing the true value (c.f. 90.6% in Experiment 1a). By comparison, the Interleaved condition ranges contained the true value 88.9% of the time and the Repeated condition 69.4%. The hit and miss rates were compared using a Cochran's Q Test, which confirmed a significant difference, $Q(2) = 9.5$, $p = .009$. McNemar's tests were used, post-hoc, to determine which conditions were driving this result. These indicated that the MOLE produced superior outcomes to the Repeated but not the Interleaved condition,

$\chi^2(1) = 8$ and 0.2, p = .002 and .327 (one-tailed). The difference between performance on the Interleaved and Repeated conditions was also significant, $\chi^2(1) = 4.45$, p = .0174 (one-tailed).

Time. Looking at Table 3, the MOLE is easily the fastest technique, 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 and suggesting that people in the two conditions examined the target stimulus for the same amount of time. It does, however, argue against the possibility that mere exposure could account for the MOLE's performance in Experiment 1a.

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, used post-hoc, 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.

Table 3. Time to complete task by condition

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

3.4 Conclusions

The results offer support for the use of repeated judgments in elicitation tasks – in line with expectations. The Repeated method, subject to the standard problems with repeated individual judgments was the worst performer. It was, however, superior to the equivalent Experiment 1a results, indicating that even making repeated judgements in situations where the participant knew they were judging the same stimulus again and again helps in improving estimates – whether due to changes in beliefs across the task or simply greater exposure time.

However, the Interleaved method, which aimed to avoid the problem of participant awareness of the repetition by locating the experimental trials within a series of distractor tasks, yielded a larger benefit (small increase in accuracy and significant decrease in overconfidence) with the same exposure time of the target stimulus. That is, ensuring the

independence of estimates increases the benefit seen from repetition – in line with previous research (Vul & Pashler, 2008).

Overall, the MOLE method was the most accurate, generated less overconfident ranges and took the least time to complete – although only on the last was its advantage over the Interleaved significant. It is also, however, generalizable to domains where the Interleaved approach is untenable (e.g., Experiments 2 and 3 described herein).

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

3.4.1 Caveats

Despite the results, there is a limitation that should be addressed. Specifically, 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.

The similarity between the MOLE and Interleaved results is also worth commenting on. Given the Interleaved process produced results nearly as good as the MOLE – in fact, statistically indistinguishable within our small sample – it is worth considering whether the ‘blind repetition’ aspect of the MOLE is the primary driver of its superiority over more basic, direct elicitation methods like those in Experiment 1a. That is, whether the other aspects (retaining uncertainty, relative judgements and washing out any anchors) are less important. To answer this, larger, more powerful studies will be required to determine whether the MOLE retains its current advantage in terms of its accuracy (i.e., the higher correlation).

4. Experiment 2: Overconfidence in Epistemic Uncertainty

4.1 Aims and Objectives

Perceptual stimuli were used in Experiments 1a and 1b to allow production of a task on which individual differences in participant knowledge would be irrelevant and which would allow use of repeated measures (and thus within-subjects designs) in a way that a more

traditional elicitation questions would not. However, given that the majority of elicitation research is undertaken on epistemic uncertainty (where the task is to recall information from memory in order to answer a question) these results could be argued to have limited generalizability. Therefore this study seeks to confirm the MOLE's benefit over alternative elicitation methods when used to elicit answers to questions where participants are relying on knowledge and memory rather than perception.

A secondary consideration for this experiment was to test whether the MOLE's use of initial, starting ranges was providing an unfair benefit – by restricting the magnitude of errors that a participant could, theoretically make.

4.2 Method

4.2.1 Participants

Participants were 60 university students and members of the general public, 27 male and 33 female with a mean age of 25.3, *(SD* = 8.9). Each received a $20 book voucher for their time. In addition, to encourage accuracy, an additional $20 voucher was promised to the best performing participant from each condition.

4.2.2 Materials and Procedure

Participants were sorted randomly into one of three conditions, coded as separate Matlab GUIs. Each presented, in a random order, the same 15 almanac-style questions with numerical answers ranging from 14.5 (% of world population living in Africa) to more than 1.7 million (area of the Australian State of Queensland in km^2). Such questions are used to create epistemic uncertainty – as participants are unlikely to know the correct answer but are likely to have some knowledge that can be used to generate a non-random estimate.

Five of these questions had answers that were percentages and, thus, had clear preset ranges (0-100) for all participants' responses. The remaining 10 questions were divided into two groups – designated *Double* and *Quintuple* according to whether the MOLE GUI used a range from zero to double the true value or zero to five times the true value as its preset range. Each of these two groups had questions from across the full range of magnitude and were selected as being of similar difficulty.

MOLE. The MOLE GUI was essentially identical to that described for Experiments 1a and 1b, except that, instead of an array of circles, participants saw a single question presented, which remained visible throughout. The only difference from the Experiment 1

method was the inclusion of a final, evaluation step where people were presented with the range calculated by the MOLE and asked to evaluate how confident they were that the true value would fall within this.

MMM. This condition (labelled MMM for minimum, maximum and most likely) was similar to the 'triangular' elicitation method from Experiment 1a except that the range was elicited prior to the best guess. This procedure was used as it gives a range estimate unaffected by the best guess but also yields a direct measure of the participant's best guess. As with the MOLE GUI, the question remained visible while all estimates were made.

Dialectical. The final condition was similar to the 'iterative' condition from Experiment 1a but drew upon Herzog and Herwig's (2009) observations regarding the use of dialectical processes in improving point estimates. For uncertainty elicitation, however, the key improvement needs to be in the range rather than the best guess and, as such, the dialectical process was used to revise the range. Specifically, after a participant made a set of estimates exactly as they would in the MMM condition, they were asked to: 1, consider the possibility that their range did not contain the true value; 2, indicate whether the true value was more likely to lie above or below their range; and, 3, revise their minimum and maximum following this thought experiment before providing their confidence estimate.

The majority of participants completed the experiment in 30 minutes or less.

4.3 Results

4.3.1 Comparisons between elicitation methods

To compare the elicitation methods, participants' confidence (that their range would contain the true value) and calibration scores (percentage of ranges containing the true answer to each question) were calculated participants. These, along with the average time taken to complete a question are summarized in Table 4.

Table 4 shows the MOLE produces the best-calibrated ranges, with participants' ranges containing the true value 72% of the time. Participants were, however, still overconfident, whether considering the expected 100% confidence interval or the evaluated confidence level, which indicated that participants expected their ranges to contain the true value ~86% of the time. That is, overconfidence is either 28% or 14%, depending on whether one uses the expected or evaluated confidence.

This was a superior result to that seen in either of the direct elicitation methods, where participants' ranges contained the true value less than 40% of the time but were predicted to

~65% of the time, which yields overconfidence scores of 60% or 25% - approximately double the bias seen in the MOLE results.

Table 4 Performance by elicitation method (means and SDs)

	Seconds/Question	Calibration	
Technique		% Hits	Confidence
MMM	39.2 (24.2)	37.3 (15.7)	67.3 (27.6)
Dial.	53.5 (28.5)	38.7 (13.8)	64.7 (28.6)
MOLE	90.3 (110.7)	72.0 (9.6)	85.7 (19.9)

* - Data is calculated from 20 participants in each group across 15 questions – thus N=300 for the time and confidence measures but N=20 for calibration as this is calculated across all 15 questions seen by an individual.

As noted in previous experiments, however, calibration is not the only measure an elicitor might be interested in. As was the case in Experiment 1, correlations between the true and estimated answers were calculated, showing a clear advantage for the MOLE, with a Spearman correlation of 0.76, CI_{95}= [.52 .91], over the Dialectical and MMM methods (*rho* = 0.07 and 0.34, respectively) – although, given a set of only 15 questions and the orders-of-magnitude differences between their answers (and errors in estimation), these are, at best, unreliable measures of the accuracy of participants' estimates.

Table 5. Absolute error by condition.

	Mean	*SD*	Median	IQR
MMM	200.4	*3444*	48.21	[16.0 - 94.2]
Dial.	1199.0	*20667*	50.38	[20.8 – 91.6]
MOLE	85.5	*101.8*	53.84	[19.5 – 88.2]

As an additional measure of the accuracy of participants estimates across the three conditions, the mean and median % absolute errors were calculated – that is: 100* |True – Estimate| / True. These values are shown in Table 5 where one sees that, in terms of their median values and interquartile ranges of the absolute error, the three techniques are largely indistinguishable - with all showing a median error of around 50%. In terms of their mean

error, however, the MOLE has a strong advantage, with far fewer extreme values skewing the results. That is, the advantage of the MOLE seems to stem from its prevention of estimates that are out by large amounts (in either direction).

4.3.2 Effect of preset ranges on elicited ranges and values

A possible objection to the previous experiments' conclusions was that the use of preset ranges might be the primary cause of the MOLE's advantage. Three different types of questions were used in this experiment to test this question; specifically, with preset ranges of 0-100 (Percentage), 0-2x the true value (Double) and 0-5x the true value (Quintuple). The expectation being that, if the MOLE's advantage lies in its use of preset ranges, then manipulating these ranges will affect it disproportionately. Specifically, one would expect no advantage for the MOLE in the percentage questions (as participants in all conditions have the same preset range) and a stronger advantage in the double questions than the quintuple questions (as the former restrict high estimates to a greater extent and has a mean - of the initial distribution of possible options - equal to the true value). Of course, the Double and Quintuple questions sets actually only differ from one another in the MOLE condition, meaning that the difference between these within the MMM and Dialectical methods are expected to be null. That is, there should be no difference between these question-types for these elicitation methods but possibly a difference in the MOLE – meaning an interaction effect would be illuminating.

Figure 8, displaying the mean confidence and calibration for each type of question and elicitation method, shows a clear advantage of the MOLE method in both confidence and calibration across all three questions types – a result confirmed by two 3x3 (condition by question-type) mixed design ANOVA with question type as the within-subjects factor, conducted for confidence and calibration.

Starting with confidence, this found significant main effects of both condition, $F(2,57)=6.9$ $p=.002$, and question type, $F(2,114)=13.0$, $p=.001$ and indicated no interaction between these, $F(2,57)=0.85$, $p = .433$. Bonferroni post-hoc tests indicated that the significant difference in condition was due to participants in the MOLE condition being more confident than those in the other two conditions, $p = .04$ and $.012$. Similarly, the effect of question type was found, post-hoc, to result from the difference between the percentage-type questions and the other two types.

The ANOVA run for calibration found significant effects of condition, question type

and the interaction between the two, $F(2,57) = 209.2$, $F(2,114) = 10.2$ and $F(4,114) = 6.2$, $p < .001$ in all cases. Post-hoc analyses confirmed that the MOLE resulted in significantly higher calibration than the other two elicitation conditions, $p \leq .001$ and that the Quintuple-type questions resulted in significantly lower calibration than the other two types, $p = .006$ and $p < .001$, for the Double- and Percentage-types respectively. This is of interest as it implies that the questions in the Quintuple set may have proved harder than those in the double set, despite people's equal confidence. Otherwise, one would expect the calibration of participants in the DIAL And MMM conditions to be equal between these question types.

Returning to Figure 8, it seems likely that the interaction effect is resulting from the unexpectedly low calibration achieved in the MOLE condition on the Percentage-type questions. In this, particular, combination of condition and question type, the degree of overconfidence in the MOLE is quite similar to that seen in the other two conditions (20% compared to 18% and 25%), which could be interpreted as being in line with the prediction the MOLE would have no advantage on questions of this type. However, the other result expected if the use of preset ranges benefits the MOLE (a greater advantage in the double than the quintuple questions) is not observed; instead, the greatest advantage of the MOLE is in the quintuple questions (12% overconfidence compared to 32% and 41%).

4.4 Conclusions

In general, the results of this experiment confirm the benefits of the MOLE procedure, despite the change from perceptual to epistemic uncertainty and from a within- to a between-subjects design. Specifically, the MOLE method resulted in both much better calibration compared to the alternative measures (~14% overconfidence compared to 28% and 30% for Dial. and MMM, respectively). By contrast, the dialectical method failed to show any significant benefit over the simple range plus best guess elicitation (MMM) – although Figure 8 suggests the dialectical method might weakly reduce confidence.

The accuracy of point estimates calculated using the MOLE method was also superior to direct estimates made by participants in the alternative elicitation methods - reducing the number of wildly wrong estimates – although it should be noted that participants found the questions hard and answers in all conditions regularly differed significantly from the truth.

Finally, the attempts to identify any role of the preset range in the advantage the MOLE enjoys were inconclusive. As noted above, the MOLE had less advantage in terms of overconfidence on the Percentage-type questions (as confidence and calibration scores were

both around 20% higher than in the other conditions, which is in line with a hypothesis holding that the preset range is responsible for the MOLE's advantage. If this were the entire story, however, one would also expect the Double–type questions to have an advantage over the Quintuple, whereas the opposite was observed. A possible confound lies in the within-subjects design, which necessitated using different questions in the Double and Quintuple conditions, with the result that they may not of equivalent difficulty. Future work could examine this more closely using the same questions with different starting range widths for the MOLE in a between-subjects design.

Even with that caveat, however, the fact that the MOLE also results in far fewer extremely low estimates also argues against the preset range being the sole cause of it superiority – particularly as regards accuracy. That is, the MOLE's preset range allows for low values just as inaccurate as the other methods but these are not observed. Finally, it should also be noted that, in terms of predictive power, 67% from 87% is a superior result to ~50% from ~70% (to understand why, consider the extreme case where 0% of ranges contain the true value when 20% are expected to). That is, given the same degree of overconfidence, we should prefer the estimates of people with higher confidence and calibration scores and, taking this into account the MOLE can, justifiably, be argued to be superior to either alternative using all three question types.

5. Experiment 3: Overconfidence in Forecasting

5.1 Aims and Objectives

This experiment compared MOLE's calibration on a forecasting task with direct elicitation wherein participants provided minimum and maximum estimates. Given that perhaps the majority of important elicitation problems involve the forecasting of future values, it was regarded as important to establish whether the advantage observed for the MOLE on perceptual and epistemic tasks remained on a forecasting task, where participants estimate ranges they are confident will contain the true value that a parameter of interest will take at a specified point in the future.

It is important to note that this design, with testing across an extended period and yet with all participants making forecasts across the same duration, results in individual results being dependent on the volatility of the parameters across that period. That is, participants using the same starting value on different days and making the same range estimate may end up with different calibration scores as a result of the true value on the target days differing. A

period of low volatility could, thus, mask poor calibration.

5.2 Method

5.2.1 Participants

Participants were 158 oil industry personnel employed in the US (*n* =115) and UK (*n* = 43). While, for confidentiality reasons, demographic data were not collected, previous work suggests a mean age of ~40 and an average of 15 years of industry experience is typical; as is a 3 or 4:1 male to female ratio (see, e.g., Welsh et al., 2006; Welsh, Bratvold, & Begg, 2005). Given the involved companies' interest in seeing results for their personnel, all participants willing to participate were accepted, rather than determining numbers in advance. However, analyses were not begun until all data collection was complete within a given location.

5.2.2 Materials

The MOLE and direct estimation methods both asked participants 10 questions regarding the values of 5 commodities/shares at times 7 and 28 days following testing. Two equivalent question sets were developed – labelled Gold and Silver after the first commodity included in each, as seen in Table 6. Some of these (e.g., oil and gas price and company share price) were selected as being directly relevant to participants' work; others as indices that industry professionals might have cause to follow for investment reasons (precious metals and stock indices); and the remainder (temperature, rainfall and windspeed at a nearby location) as variables that any local person could make a reasonable attempt at forecasting.

For the US participants, the quiz questions were coded into a graphical user interface (GUI) for delivery via the MOLE but delivered as a paper and pencil test for the direct estimation. For the UK participants, both the MOLE and direct estimation methods were delivered via GUI. Figure 9 shows the GUI as it appears during elicitation using MOLE, with the inclusion (for the first time) of an 'Unselect' button that allowed participants to change the option they had selected in cases where they had accidentally pressed the wrong button.

5.2.3 Procedure

Participants were tested in small groups (2-4) at company offices over a period of approximately 1 month – in each country. Which quiz a participant undertook under each elicitation method was determined randomly. That is, approximately half of participants completed the Gold quiz using the MOLE and Silver using the standard elicitation, while the

remainder did the reverse. Which of the methods was delivered first was also randomized.

Standard Elicitation Procedure

Under the standard elicitation condition, participants were asked to give ranges they were *certain* would contain the true value of the parameters of interest at the specified time. That is, they were asked for minimum and maximum values.

These were either recorded on a paper copy of the quiz or entered directly into the GUI. Prior to testing, participants were asked to record the *current* value of the parameter of interest – to ensure that they had some idea of what the true value was and thus better reflect real forecasting tasks where people forecast values that they are familiar with.

It was decided not to ask participants for a best guess as this affects the width of elicited ranges in complex ways (see, e.g., Block & Harper, 1991; Heywood-Smith et al., 2008), including the suggestion that it affects ranges differentially according to a person's level of expertise (Bruza et al., 2011).

Table 6. Commodities/parameters by quiz.

Q.	Forecast Window	Quiz 1 (Gold)	Quiz 2 (Silver)
1	7 days	Gold price	Silver price
2	28 days	Gold price	Silver price
3	7 days	Maximum Temp	Minimum Temp
4	28 days	Maximum Temp	Minimum Temp
5	7 days	Rainfall total	Wind Speed
6	28 days	Rainfall total	Wind Speed
7	7 days	Share price	Share index
8	28 days	Share price	Share index
9	7 days	Oil price	Gas price
10	28 days	Oil price	Gas price

NB – the specific values asked from varied across locations. E.g., the Share price asked for was for each participant's own company and the share index was for their country of residence (Dow Jones for US; FTSE100 for UK).

MOLE Procedure

As no true values existed at the time of the experiment, the MOLE required the experimenter to set initial bounds on the range of values that the computer would use – based on extrapolations of historical data or natural limits (where available). The bounds used for the different quiz questions are shown in Table 7. Note that some were based on the

parameter's current value while others were based on historical data. In both cases, however, the participant was tasked with entering the current value into the MOLE GUI immediately prior to the elicitation beginning. In this way, participants were assured of knowing something about the parameter in question. The only other difference from previous versions of the MOLE was that the participants rated their confidence in their choice on a verbal scale from guessing to very high[1] (as seen in Figure 9) rather than 50-100%.

Given the use of a simple, range elicitation as the comparison condition, best guess values were not calculated for this experiment. Participants were not made aware of the underlying MOLE algorithm or its starting ranges, ensuring that attempts to 'game the system' would be made blind.

Table 7. Initial bounds for MOLE process.

	US		UK	
Q.	Gold	Silver	Gold	Silver
1	±5%	±5%	±10%	±10%
2	±10%	±10%	±10%	±10%
3	30-110F	30-110F	-20-40C	-20-40C
4	30-100F	30-110F	-20-40C	-20-40C
5	0-7 in.	0-60 mph	0-100mm	0-90kmph
6	0-20 in.	0-60 mph	0-200mm	0-90kmph
7	±5%	±5%	±5%	±5%
8	±10%	±10%	±10%	±10%
9	±5%	±10%	±5%	±10%
10	±10%	±20%	±10%	±20%

Note: ±% indicates bounds were calculated from the current value of the parameter. Note 2: the UK 7-day bounds are, in places, wider than their US equivalents as detailed below.

5.3 Results

5.3.1 Methodological Concerns

The US sample was collected several months before the UK sample and, as such, observations from this were used to update our process for determining bounds. Specifically,

[1] This scale was mapped over the top of the 50% - 100% confidence scale used in previous versions of the MOLE – as a result of discussions with the companies providing participants. While this, necessarily, reduces our ability to interpret results, it should be noted that the effect of this can only be to narrow ranges when the numerical scale might otherwise leave it intact. Thus, this change can only hinder the MOLE.

it was observed that the bounds used for the Silver price underestimated the volatility in the market – preventing a number of participants from being able to capture the true value in their final ranges, no matter what choices they made during the MOLE. In light of this, ranges used for the UK sample were widened on this question and analyses exclude this question from the US dataset. Otherwise, the differences in bounds reflect differences in expected weather for the participants' local areas and changes of units from metric to imperial where appropriate.

Another methodological concern was the possibility that the change from a numerical to verbal labels on the MOLE GUI's confidence slider might negatively impact the MOLE's performance. As noted below, however, the calibration observed in this experiment was very similar to that achieved by the MOLE in previous experiments. As such, this change appears not to have any significant effect and is not discussed further.

5.3.2 Equivalency of Quizzes

Given the differences in questions answered by the two samples, described above, individual analyses (i.e., t-tests) were used within each sample in preference to a 2x2 ANOVA comparing the groups and time-frames simultaneously. Apart from the effect noted above for the silver question, the US sample's performance on the questions from the Gold and Silver quizzes was statistically equivalent. Calibration on the 'Gold' and 'Silver' question sets was compared for both 7 day and 28 forecasts using Welch's t-tests. These showed no difference between people's calibration on the two sets of questions, $M = 82.8$ and 84.0, $t(228) = 0.42$, $p = 0.674$ on the 7 day forecasts and $M = 84.0$ and 85.7, $t(228) = 0.58$, $p = 0.566$ on the 28 day forecasts.

The UK sample is slightly more complex in that, while there is no observed difference between participants' performance on the Gold and Silver quizzes under the MOLE, there is one using the standard elicitation method, with the average calibration being 20% lower on the Gold quiz. On examination of the data, it was noted that, during the period of testing for the UK sample, the parameters on the Gold quiz happened to be markedly more variable than those on the Silver quiz. The average difference between the minimum and maximum values observed for the various parameters across the date range (i.e., $D = (Max-Min)/Max$) was 0.37 for the Gold quiz compared to 0.22 for the Silver.

In light of the larger US sample's results, however, it was decided that this did not call into question the equivalency of the questions, per se.

5.3.3 Calibration

Given differences between the UK and US samples question sets and differences in starting commodity values (and thus MOLE starting ranges) across the duration of the experiment, the actual widths of ranges are not directly comparable in this experiment. Therefore, analyses focus on calibration: calculated as the proportion of ranges containing the true value (given that 100% confidence intervals were elicited). Figures 10 and 11 show mean calibration by forecast window and elicitation condition for the US and UK samples, respectively.

Looking at Figure 10, one sees two very clear results. The first is that the forecast length did not affect calibration – with little difference seen between the 7 and 28 day forecasts under either condition in paired samples t-tests, $t(114) = 0.493$ and 1.81, $p = .623$ and .073, $A = .526$ and .539 (common language effect size - specifically, the measure of stochastic superiority; Vargha & Delaney, 2000), for the direct estimation and MOLE conditions respectively. That is, while participants did, in both conditions, increase the width of their ranges for the 28 day forecasts relative to the 7, the benefit in terms of calibration was zero as the additional width was offset by the parameters' greater volatility in the longer term.

The second observation is that the MOLE method produced markedly better calibration for both 7 and 28 day forecasts – with approximately 17% more of its ranges containing the true value than the direct estimation method. Paired sample t-tests comparing participants' calibration on the two elicitation methods (for each forecast length separately) unambiguously support this, $t(114) = 6.92$, $p << .001$ for the 7 day data and $t(114) = 6.06$, $p << .001$ for the 28 day forecasts. The effect size was large in both cases, $A = 0.734$ and 0.730.

Turning to Figure 11, one sees a similar pattern of results – although the 28 day result for the direct estimation method shows a decline in calibration as a result of the greater volatility in the Gold quiz questions discussed above. A paired sample t-test indicated that the difference observed here was significant, $t(42) = 3.1$, $p = .004$, $A = .604$. A second, paired sample t-test indicated no difference between participant's 7 and 28 day forecast calibration using the MOLE, $t(42) = 0.22$, $p = 0.824$, $A = 0.521$.

The difference between participants' mean calibration on the MOLE and direct estimation was 17% on the 7 day forecast and 27% at 28 days. Paired sample t-tests comparing mean calibration at each forecast length confirmed these differences were significant, $t(42) = 4.3$ and 5.9, $p < .001$ and $p << .001$, $A = 0.734$ and 0.779.

Looking at Figure 11 and the t-test described above, it seems clear there is an interaction effect – with the longer period affecting calibration only for participants in the direct estimation condition. That is, greater volatility on the Gold quiz questions (discussed above) decreased calibration for participants undertaking the direct estimation conditions, but not for those answering the same questions using the MOLE.

5.4 Conclusions

The results confirm that the MOLE's advantaged over direct estimation elicitation methods in Experiments 1a, 1b and 2 – extending from perceptual and epistemic paradigms to a forecasting approach with greater applicability to real world problems.

While the MOLE does not eliminate overconfidence (this may, in fact, be impossible where error is involved - as discussed by Soll & Klayman, 2004), it reduces it markedly compared to direct estimation approaches to elicitation. Overconfidence when using the MOLE is around 7% in Experiment 3 – about one third of the direct estimation overconfidence of around 25%.

Some results do, however, require additional explanation; for instance, in the UK sample, additional volatility in some parameters across the experiment's (moving) forecast window led to a marked decrease in calibration in the direct estimation task but not the MOLE. A likely cause of this is the outside-in method the MOLE uses to construct its final range. As shown in Figure 1, this is predicted to result in wider ranges – as observed – but also to have a greater effect where uncertainty is higher, as would be expected in the longer forecast window.

By requiring participants to definitively rule values out before removing them from consideration (rather than asking whether they should be included), the MOLE preserves as much of a person's 'region of uncertainty' as possible. Given the participant (presumably) believes any value within this range is possible – all of them should fall within a 100% confidence interval and the MOLE makes this far more likely.

That this makes ranges wider is unsurprising but the fact that it also prevents the drop off in calibration seen with unexpectedly high volatility demonstrates the approach's strength and seems to have parallels with Yaniv and Foster's (1995) accuracy/informativeness trade-off. That is, people accept values presented by the MOLE as possible, despite the fact that they would not report such values themselves for fear of them being deemed uninformative.

Another interesting observation is that participants maintain the same calibration

when predicting further into the future by giving wider ranges, mirroring the observation that expert and novice forecasters have similar levels of overconfidence despite differences in knowledge (McKenzie, Liersch, & Yaniv, 2008). This suggests people may have a stable, preferred levels of calibration or informativeness (an idea supported by findings relating overconfidence and need for cognitive closure; for details, see: Kaesler, Welsh, & Semmler, 2016)

5.4.1 Caveats

As noted above, both the MOLE and direct estimation conditions are assumed to yield 100% confidence intervals – that is intervals the participant believes will definitely contain the true range. While this could, in the direct estimation condition, lead to 'sandbagging' (i.e., generating 0 to ∞ ranges to guarantee success), this is not observed in the data due to people's tendency towards informativeness (Yaniv & Foster, 1995). (In fact, such wide ranges are not generally appropriate. For example, "temperature measured at Heathrow Airport" will not ever exceed 400°C - the autoignition point of jet fuel and, thus, the temperature at which the airport (and its thermometers) will cease to exist.)

It should also be noted that a typical calibration task asking for 80% confidence intervals can equally easily be 'gamed' by providing 80% extremely wide ranges and 20% extremely narrow (or just plain wrong) estimates. Any tendency that a person has towards such behaviour would, presumably, benefit their calibration scores in the direct estimation task to a greater extent than in the MOLE which, as noted above, did not make clear to participants the process by which it created a range from their responses. Thus, to the extent that such effects impact the data, it would be expected to erode differences between the two conditions – which remain marked. Future work could, however, benefit from a consideration of proper scoring rules (Brier scores or others; for details see, e.g., Carvalho, 2016) which penalise such attempts to manipulate calibration.

The second concern is the requirement that the experimenter set the initial bounds for the MOLE – as demonstrated by the authors' own failure to account for the volatility of the silver price for the US sample. While this increases the potential for overconfidence in the MOLE results – by causing cases where it is impossible to create a range that contains the true value – more judicious use of historical data and natural bounds renders this a relatively minor concern. Certainly, defining an initial range is a problem shared with any elicitation method that seeks to guide participants to consider a wider range (see, e.g., Haran, Moore, &

Morewedge, 2010, who ask participants to assign probabilitites across the full range of possible answers - as defined by the experimenters).

6. Discussion

The results of the four experiments described herein paint a consistent picture, with the MOLE process outperforming all of the elicitation methods to which it was compared – whether being used to elicit perceptual estimates, epistemic uncertainty or forecasts. This is a convincing demonstration of the potential benefits of designing elicitation tools in line with what we already know about human cognition and how it affects values elicited from people.

Key points of interest from the four experiments are as follows. Experiment 1a demonstrated that the MOLE, in addition to improving calibration on a task where performance was otherwise poor, markedly improved the accuracy of estimates. Examination of individual responses also demonstrated that participants' final best guesses (as calculated by the MOLE) were not being anchored by the first options shown to them, providing preliminary support for the idea that the MOLE's presentation of multiple values could erode anchoring effects.

Experiment 1b suggested that this improvement in accuracy could not be explained simply by the increased exposure of the stimulus and, instead, that repeated measurement of the same parameter explained much of the MOLE's benefit – although, as was explained at the time, how amenable alternative repeated measures methods (like the Interleaved elicitation described herein) are to non-perceptual tasks is arguable. Experiment 2 showed that the MOLE maintained an advantage over direct estimation elicitation methods when changing from a within to between-subjects design and in elicitation tasks involving epistemic rather than perceptual uncertainty. It also demonstrated that the benefit of the MOLE can not simply be attributed to a benefit provided by the preset, starting ranges required by the MOLE. Finally, Experiment 3 demonstrated that the MOLE produces better forecast ranges than direct estimation of such ranges; and that the MOLE better protected against periods of high volatility in the parameter values – presumably as a result of its tendency to preserve more of an elicitee's uncertainty.

6.1 Using the MOLE

The MOLE process, in its current form, has been shown to be a better method for eliciting range estimates from people than any of the direct estimation variants to which we

compared it. The MOLE process is, however, also amenable to variation for more specific purposes. For instance, the MOLE's method for determining the range generates a 100% confidence interval, whereas some technical uses for elicited values require an 80% confidence interval (10^{th} and 90^{th} percentiles, for example). These can, however, be generated from the MOLE's outputs if one makes some basic assumptions. Given that the MOLE produces a minimum, maximum and best guess, it is a simple matter to fit a triangular distribution to these points and, from this, any desired fractile or percentile can be extracted. Should another distribution be desired, these too can be fitted - using a participant's confidence ratings (which contain information about the relative likelihood with which the person believes the true value will fall into different regions of the total, feasible range) to generate appropriate parameters for a normal or beta distribution – from which the desired percentiles can, again, be generated.

An alternative use, for situations where fitting a distribution might pose difficulties (bimodal distributions, for instance), is to use the MOLE to generate the feasible range and then use this as the basis for a secondary elicitation process – such as Haran et al's (2010) SPIES procedure, which requires an elicitor to lead an elicitee through a process of assigning likelihoods to all possible values of a parameter – essentially building a subjective probability density function by hand. The use of the MOLE as a precursor task would limit the elicitation to those regions that the participant considered feasible – saving time and effort.

Another question regards how a person should define the MOLE's starting range. As noted earlier, for questions with natural limits (e.g., percentages), we recommend using those as the starting range. In the absence of these, starting with very wide ranges is preferable as the MOLE process allows the user to swiftly remove areas that they do not view as feasible but has no mechanism for adding range if areas that the user would consider feasible are precluded from the starting range. Where databases of prior outcomes exists, these can be used as starting points but the MOLE's starting range should also allow for values lying outside the currently observed range. Ideally, of course, the MOLE starting ranges should also be set by someone other than the intended user - or at least well in advance - so as to prevent their affecting the elicited responses.

6.2 Caveats and Future Research

6.2.1 Future Research

Additional work is required to determine whether the current mechanism for reducing

the MOLE's bounds is too conservative or, even, not conservative enough. That is, whether people are accidentally removing sections of range that they do not intend to or unable to remove sections that they consider unfeasible. The current MOLE process does not have a mechanism for examining this but it could be tested experimentally; for example, by occasionally providing values from outside the current range as a test that they are, in fact, considered infeasible.

A second line of questioning relates to the optimal number of iterations that the MOLE should run for. As noted throughout, the MOLE should generate a 100% confidence interval by allowing participants to cut away portions of the range they do not consider feasible. Whether the 10 iterations used in our experiments is sufficient for these purposes will, of course, be context dependent. Where starting ranges are particularly wide, it may require more iterations for a person to finish cutting their range. That said, in our experiments, that fact that participants' 'evaluated confidence' in MOLE ranges was less than 100% indicates that our participants regarded the MOLE as having cut the range sufficiently – that is, they were not being left with ranges so wide that they could have been cut further while still being regarded as 100% confidence intervals.

Given that the MOLE procedure is designed to improve accuracy as well as calibration – via repeated judgements and the elimination or watering down of anchoring/priming effects - altering the number of iterations the MOLE runs for and observing the effect this has on best estimates would be another valuable extension.

This work would seem to lead, naturally, to consideration of the best algorithms for selecting values to be presented to participants. Currently, the MOLE selects values randomly from a uniform distribution covering the remaining range at any point in the experiment and runs for a set number of iterations. A more intelligent algorithm, however, could take into account past values or select the most efficient comparisons for testing a participant's range or determining when the process should be terminated.

Finally, while the MOLE has been designed with the reduction of anchoring effects in mind and the preliminary test in Experiment 1a supports this idea, there is a clear need to expand on this work with the deliberate introduction of anchoring values and observation of how these impact on estimates at various stages of the MOLE process. More generally, there is a need to tease apart the impacts of the four underlying assumptions of the MOLE to see which of retention of uncertainty, relative judgements, repeated judgement and mitigating against anchoring are driving the MOLE's results.

6.3 General Conclusions

The MOLE produces ranges significantly wider than those generated by participants required to directly estimate the minimum and maximum points of a range, resulting in markedly less overconfidence. This holds true across a range of alternative elicitation methods and across three distinct domains – perceptual, epistemic and forecasting tasks.

Given the common observation that people, in general, are affected by overconfidence, the use of elicitation tools such as the MOLE, designed in accordance with established psychological theory and relying on cognitive abilities people are comfortable using, seems a useful method for improving the accuracy of range estimates.

7. 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.

Bruza, B., Welsh, M. B., Navarro, D. J., & Begg, S. H. (2011). *Does anchoring cause overconfidence only in experts?* Paper presented at the Annual Meeting of the Cognitive Science Society (33rd: 2011: Boston, USA) CogSci 2011.

Carvalho, A. (2016). An Overview of Applications of Proper Scoring Rules. *Decision Analysis, 13*(4), 223-242.

Ferretti, V., Guney, S., Montibeller, G., & von Winterfeldt, D. (2016). *Testing best practices to reduce the overconfidence bias in multi-criteria decision analysis.* Paper presented at the System Sciences (HICSS), 2016 49th Hawaii International Conference on.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*(1), 35-42.

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

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making, 5*(7), 467-476.

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.

Heywood-Smith, A., Welsh, M. B., & Begg, S. H. (2008). *Cognitive errors in estimation: does anchoring cause overconfidence?* Paper presented at the Society of Petroleum Engineers 84th Annual Technical Conference and Exhibition, Denver, Colorado.

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.

Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: a naive sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678.

Kaesler, M., Welsh, M. B., & Semmler, C. (2016). Predicting overprecision in range estimation. . In A. Papafragou, Grodner, D., Mirman, D., & Trueswell, J.C. (Ed.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus, Giroux.

Kahneman, D., Lovallo, D., & Sibony, O. (2011). Before you make that big decision. *Harvard business review, 89*(6), 50-60.

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.

McKenzie, C. R., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes, 107*(2), 179-191.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and Motivational Biases in Decision and Risk Analysis. *Risk Analysis, 35*(7), 1230-1251.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.

Mussweiler, T. (2002). The malleability of anchoring effects. *Experimental psychology, 49*(1), 67.

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26*(9), 1142-1150.

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

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs and real estate: an anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*, 84-97.

Piatelli-Palmarini, M. (1994). *Inevitable Illusions: How mistakes of reason rule our minds*. New York, NY: John Wiley & Sons.

Russo, E. J., & Schoemaker, P. J. H. (1992). Managing Overconfidence. *Sloan Management Review, 33*, 7-17.

Russo, E. J., & Schoemaker, P. J. H. (2002). *Winning Decisions* (1st ed.). New York: Currency Doubleday.

Soll, J. B., & Klayman, J. (2004). Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*(2), 299-314.

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

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

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

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*(2), 101-132.

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

Welsh, M. B., & Begg, S. (2015). Reducing overconfidence in forecasting with repeated judgement elicitation. . *Proceedings of the Annual Cognitive Science Society Meeting., 37*, 2637-2642.

Welsh, M. B., & Begg, S. H. (2016). What have we learnt? Insights from a decade of bias research. *APPEA Journal, 56*, 435-450.

Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2006). *SPE 102188: Correcting common errors in probabilistic evaluations: efficacy of debiasing.* Paper presented at the Society of Petroleum Engineers 82nd Annual Technical Conference and Exhibition., Dallas, Texas, USA.

Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2007). SPE 110765: Modeling the economic impact of cognitive biases on oil and gas decisions. *Proceedings of the Society of Petroleum Engineers 83rd Annual Technical Conference and Exhibition*.

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*.

Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005). SPE 96423 - Cognitive biases in the petroleum industry: impact and remediation. *Proceedings of the Society of Petroleum Engineers 81st Annual Technical Conference and Exhibition*.

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.), *30th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Welsh, M. B., Lee, M. D., & Begg, S. H. (2009). Repeated judgments in elicitation tasks: efficacy of the MOLE method. In N. Taatgen, H. v. Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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.

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.

Yaniv, I., & Foster, D. D. (1995). Graininess of judgment under uncertainty: an accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, 124*(4), 424-432.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making, 10*(1), 21-32.

Figure Captions

Figure 1. Pictorial representation of estimating the low-end value of an uncertainty range, working from: (a) the best guess; (b) a minimum value. Note that working from best guess rather than the minimum value results in a higher low-end estimate and thus a narrower range overall due to the width of the person's region of uncertainty regarding the low-end value.

Figure 2. Experiment 1a MOLE GUI.

Figure 3. Mean calibration by elicitation method (Exp. 1a).

Figure 4. Scatterplots of estimated and actual number of circles across all participants and trials by condition.

Figure 5. Histograms of correlations between individual participant's estimates and true values under each elicitation condition.

Figure 6 Experiment 1b. MOLE GUI.

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

Figure 8 Self-rated confidence in final range and calibration of participants by question type and elicitation process.

Figure 9. GUI showing snapshot of MOLE forecasting process. The participant has made their selection and is being asked how confident they are that their selected value is closer to the true value.

Figure 10. Mean calibration by elicitation condition and forecast window (US sample).

Figure 11. Mean calibration by elicitation condition and forecast window (UK sample).

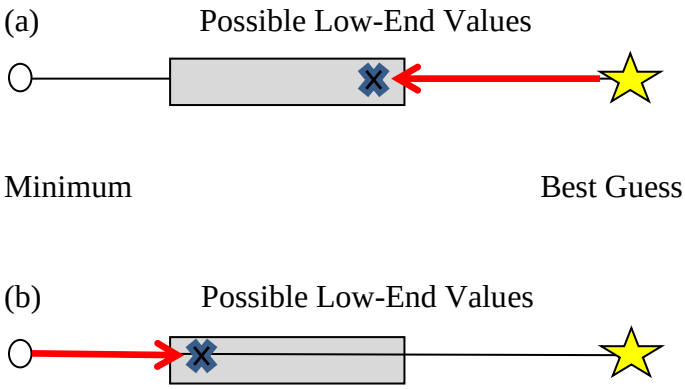

Figure 1. Pictorial representation of estimating the low-end value of an uncertainty range, working from: (a) the best guess; (b) a minimum value. Note that working from best guess rather than the minimum value results in a higher low-end estimate and thus a narrower range overall due to the width of the person's region of uncertainty regarding the low-end value.

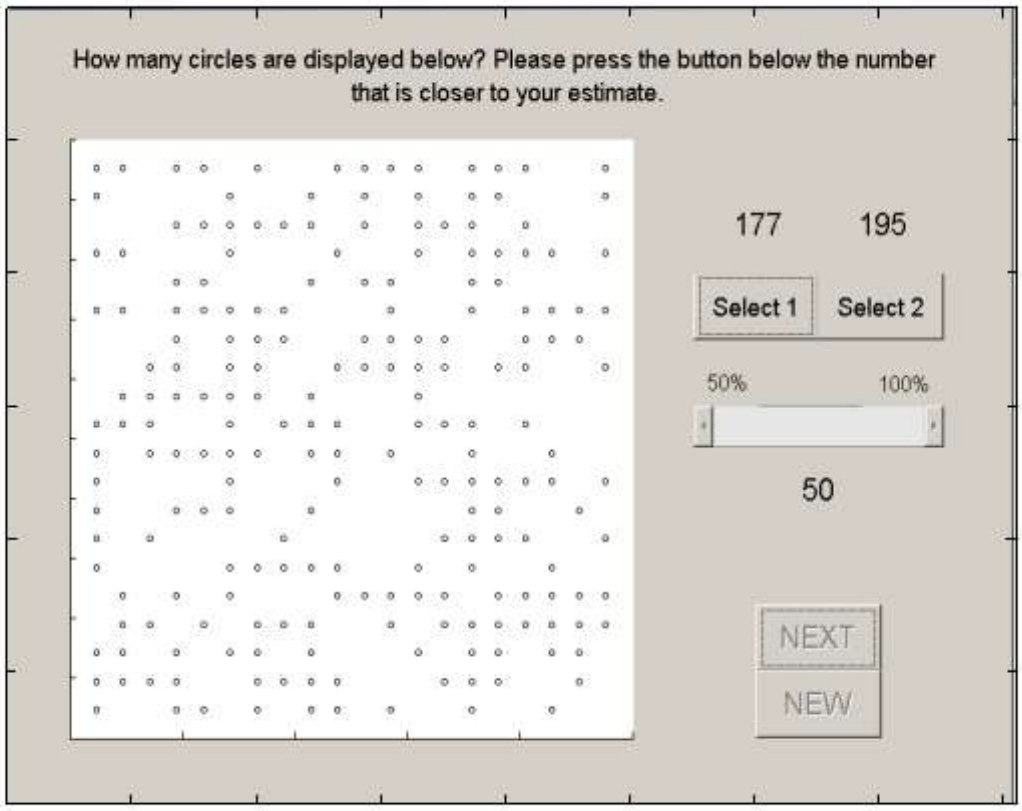

Figure 2. Experiment 1a MOLE GUI

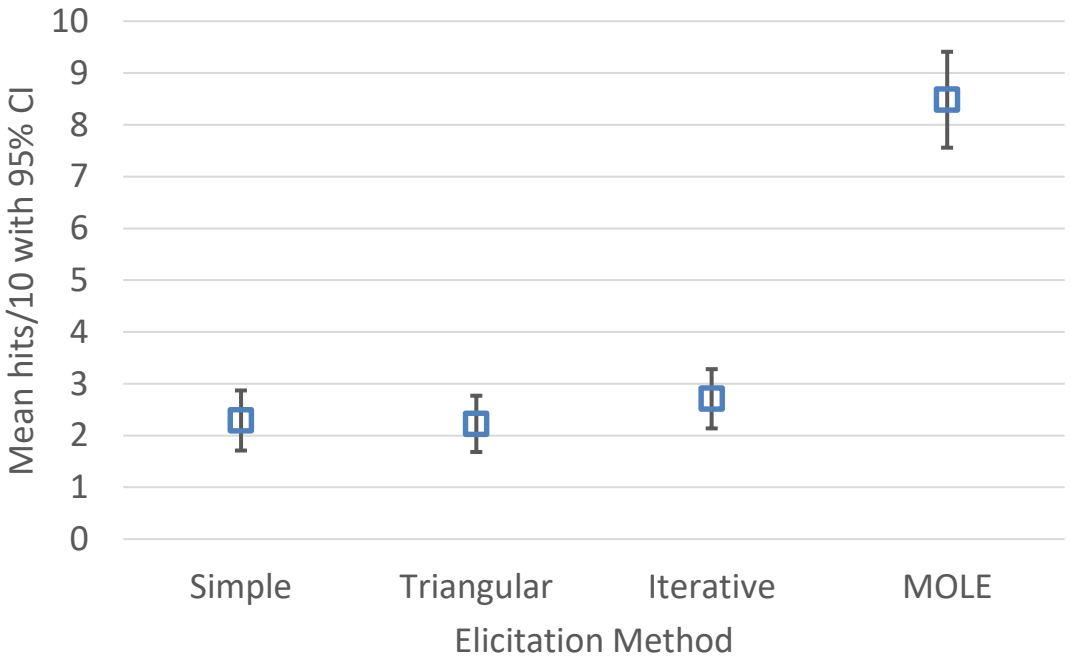

Figure 3. Mean calibration by elicitation method (Exp. 1a)

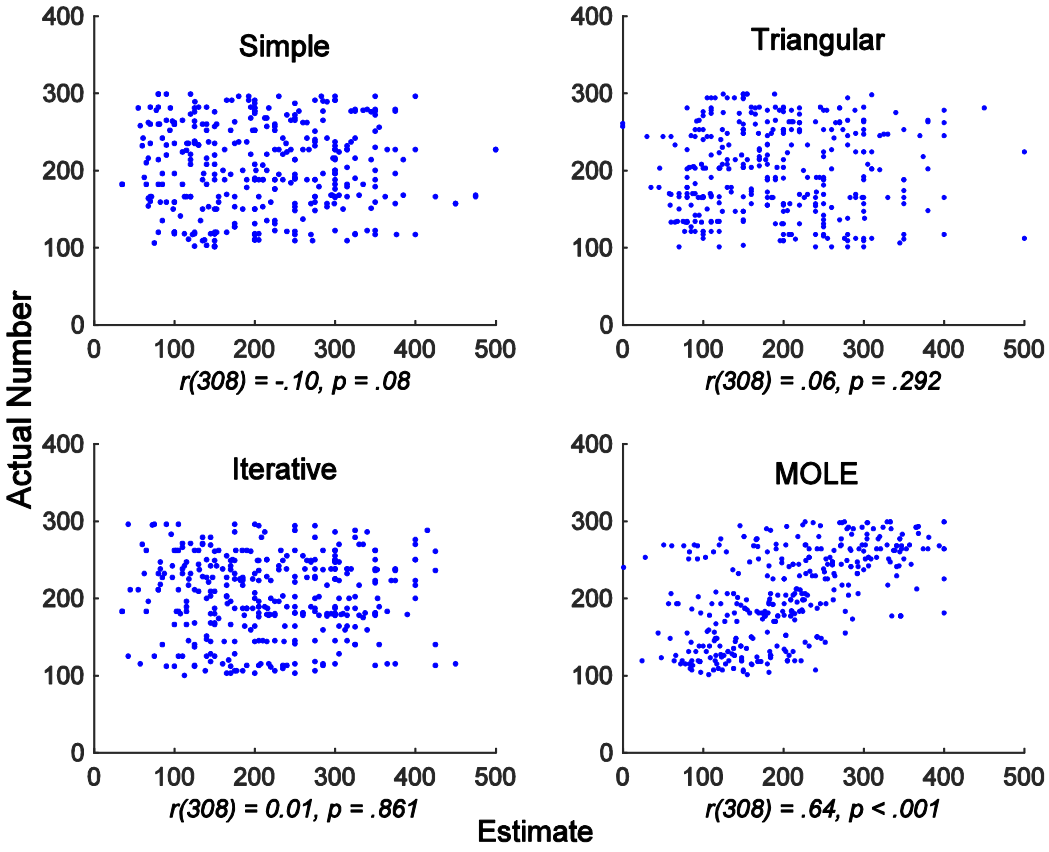

Figure 4. Scatterplots of estimated and actual number of circles across all participants and trials by condition.

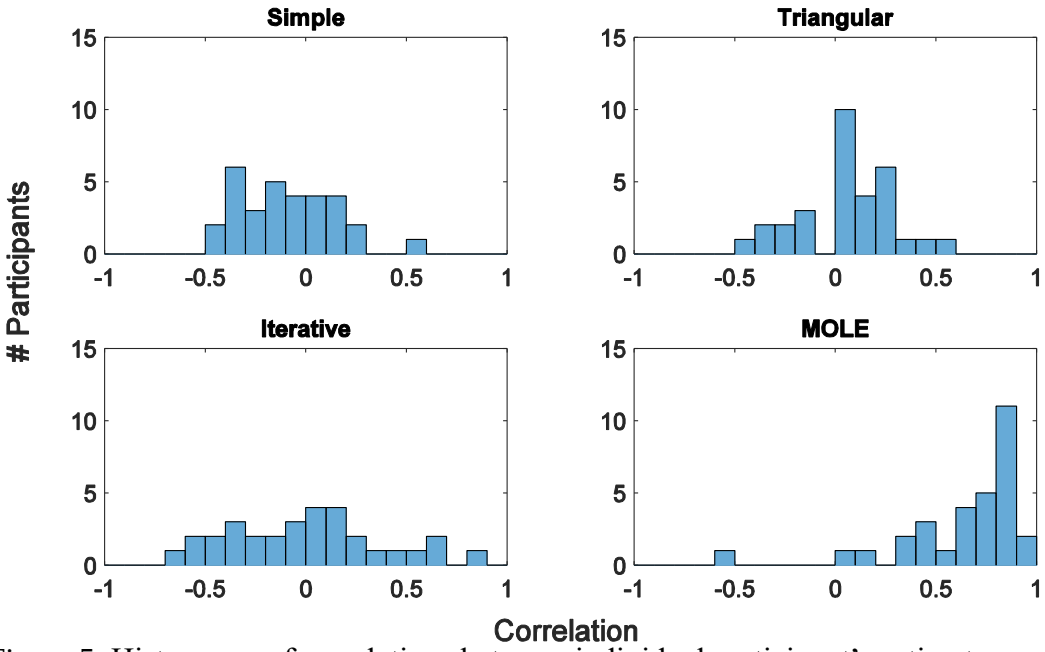

Figure 5. Histograms of correlations between individual participant's estimates and true values under each elicitation condition.

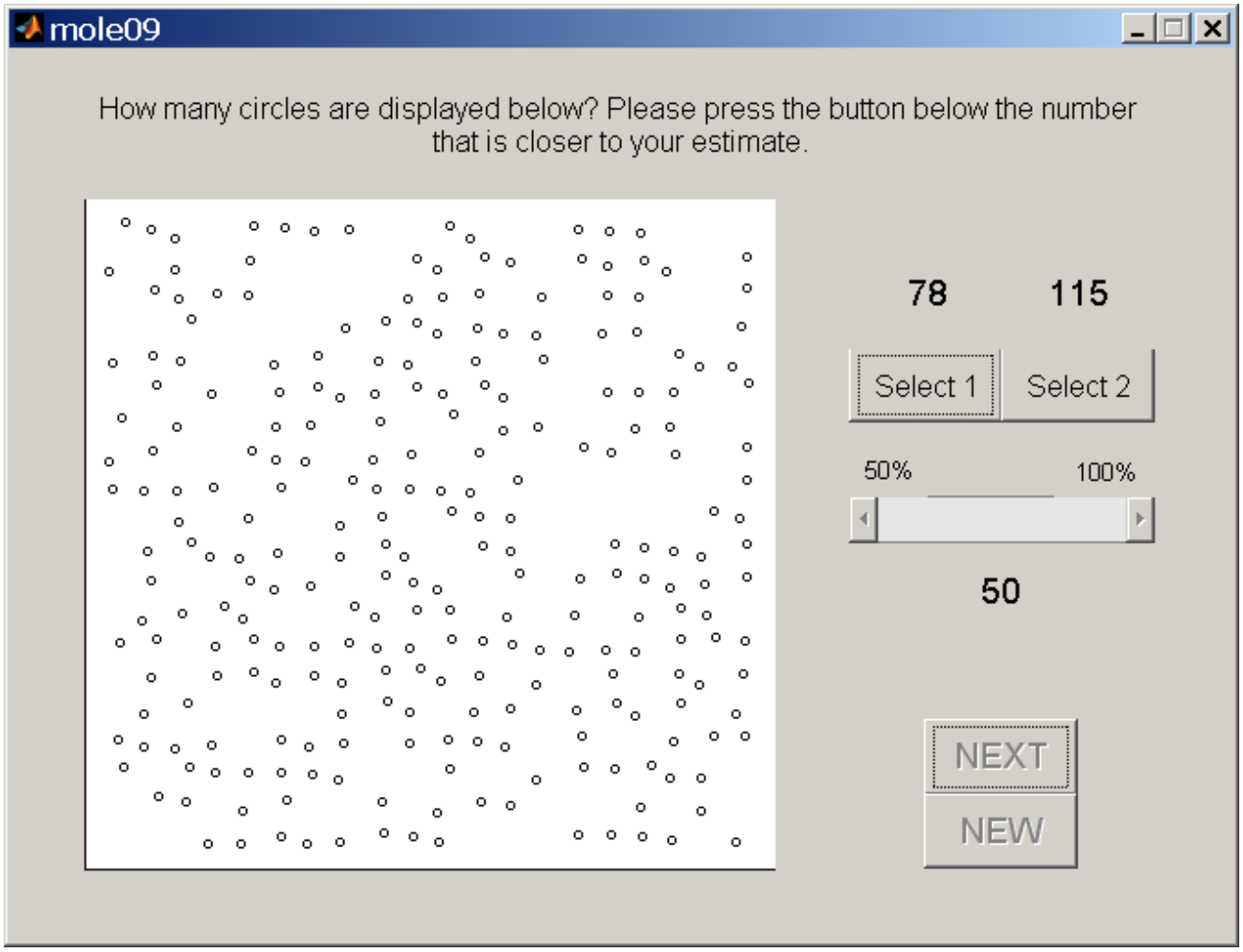

Figure 6. Experiment 1b. MOLE GUI.

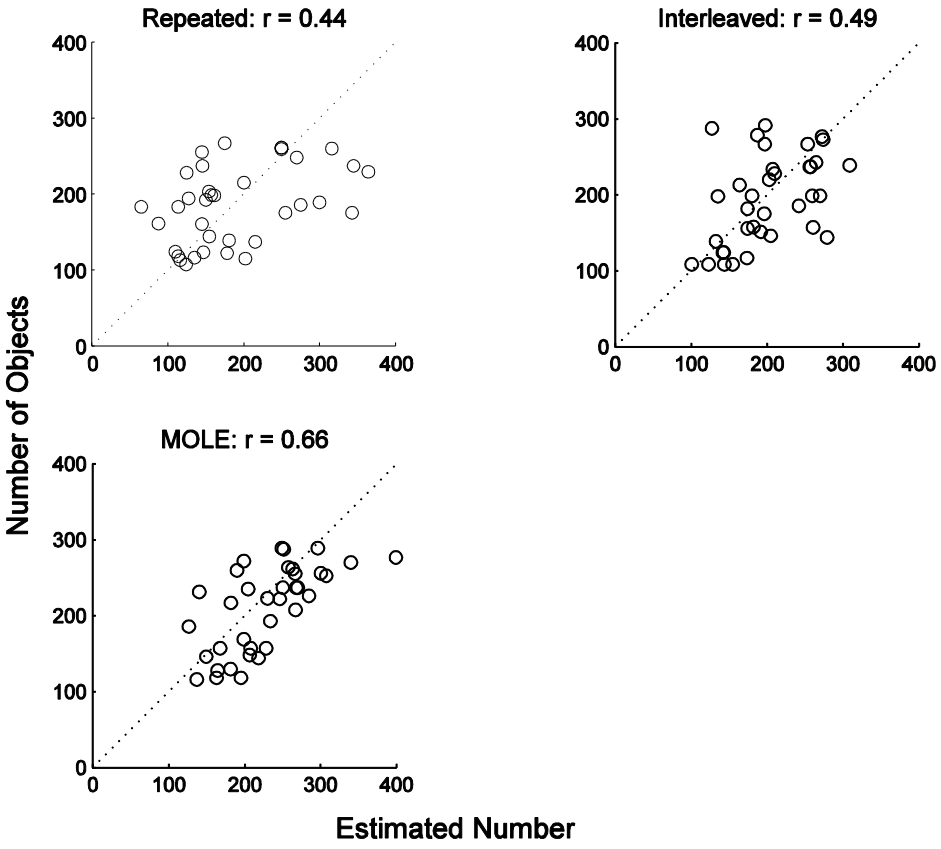

Figure 7. Scatterplots of true and estimated number of circles in arrays. N = 36 in all cases.

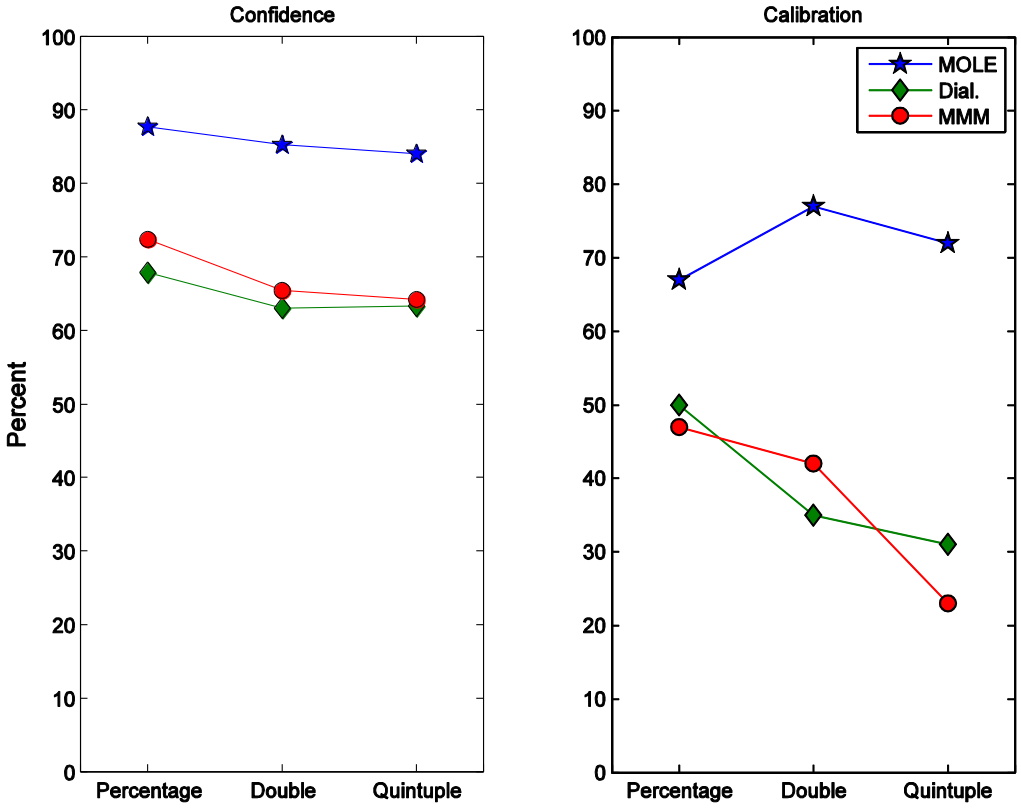

Figure 8. Self-rated confidence in final range and calibration of participants by question type and elicitation process.

Figure 9. GUI showing snapshot of MOLE forecasting process. The participant has made their selection and is being asked how confident they are that their selected value is closer to the true value.

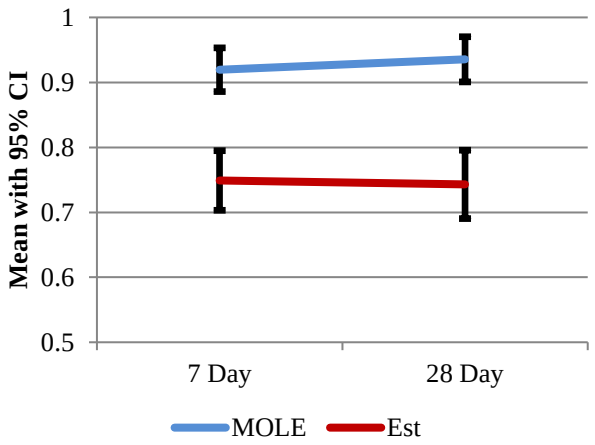

Figure 10. Mean calibration by elicitation condition and forecast window (US sample)

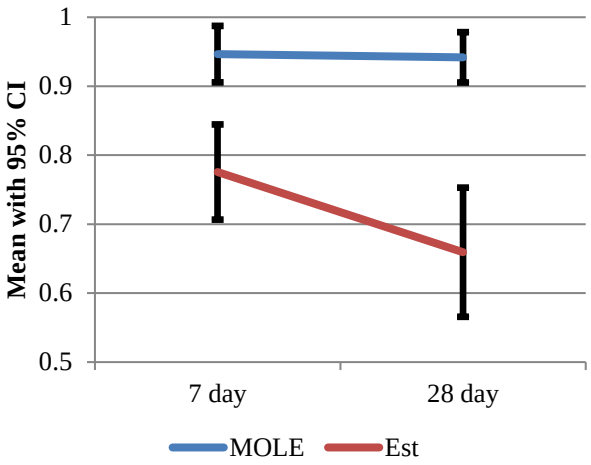

Figure 11. Mean calibration by elicitation condition and forecast window (UK sample)