Overconfidence and the MOLE:
Investigating the role of anchoring and individual differences

[redacted]

This thesis is submitted in partial fulfilment of
the Honours degree of Bachelor of Psychological Science

School of Psychology
The University of Adelaide
October 2017

Word Count: 11,395

Table of Contents

List of Figures

List of Tables

Abstract

Overconfidence bias results in the production of ranges that fail to include the true value as often as confidence levels would dictate. More or Less Elicitation (MOLE) is a tool that has been demonstrated to reduce overconfidence, through improved accuracy and calibration. This study investigated MOLE performance and tested a prior assumption that one basis for the MOLE's success is due to its ability to overcome anchoring. The MOLE was found to improve calibration, but not accuracy; and while the MOLE avoided the effects of anchoring, it was found that anchoring was unrelated to overconfidence. The study also confirmed narrow range widths, elicitation format and the Informativeness-Accuracy Trade-off (IAT) as causes of overconfidence and demonstrated that the MOLE addressed all of these factors. The study also investigated three individual difference measures for predicted or previously observed links to performance on elicitation tasks: Need for Closure (NFC), openness and conscientiousness. Conscientiousness did not improve MOLE performance, in contrast to predictions; and openness was not related to anchoring, contrary to previous findings. NFC was somewhat related to range widths, however further investigation is required. Knowledge of the subject matter was the most compelling factor, with increased knowledge relating to improved performance in many regards.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and, to the best of my knowledge, this thesis contains no materials previously published except where due reference is made. I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying. I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the school to restrict access for a period of time.

Acknowledgments

I am extremely grateful to my family for their support and patience while I completed my thesis. My friends who were also completing honours were encouraging and great sounding boards. Finally, I would like to thank my supervisor, Dr. Matthew Welsh, who guided me through this process. I appreciate Matthew’s support, humour and generosity with his time.

1 Overconfidence

1.1 What is Overconfidence?

In the face of uncertainty, we rely on people's estimates to guide our decisions; such estimation is useful, but it is also prone to systematic errors, known as biases (Kahneman, Slovic, & Tversky, 1982). One such bias is *overconfidence*, which is succinctly defined as an "excessive certainty that one knows the truth" (Ferretti, Montibeller, Guney, & von Winterfeldt, 2016, p. 1548).

Herein, 'overconfidence' refers more specifically to the form of overconfidence that Moore and Healy (2008) define as *overprecision*. This aspect of overconfidence relates to range estimation, a form of judgement that is prevalent in decision analysis (Ferretti et al., 2016). In range estimation tasks, people produce upper and lower bounds for potential values of an unknown quantity. Overprecision describes the tendency for people to produce ranges that do not contain the true value as often as they believe they will. For example, when people are asked to provide a range within which they are 90% sure that a true value will fall, the true value actually falls within the range less often than 90%; in fact, 80-90% confidence intervals are often found to contain the true value less than 50% of the time (Alpert & Raiffa, 1982; Lichtenstein, Fischhoff, & Phillips, 1982; Soll & Klayman, 2004).

1.2 Implications of Overconfidence

The overconfidence bias has implications for industries that rely on people's estimates of future quantities. In the oil and gas industry, 80% confidence ranges contain the true value less than 50% of the time (Hawkins, Coopersmith, & Cunningham, 2002); the result of these errors can be losses of tens of billions of dollars (Welsh & Begg, 2016). Overconfidence also has implications in other industries: Moore and Healy (2008) suggest it has been the cause of war and stock market bubbles; Ferretti et al. (2016) state that overconfidence can have negative impacts on financial trading activity and decisions made by health professionals, leading to mistreatment.

1.3 The Cause of Overconfidence

Overconfidence has been the basis of many studies and the effect has been shown to be very robust. However, the underlying causes are not well understood (Gigerenzer, 2000; Soll & Klayman, 2004). Indeed, the very ubiquity of the bias makes it difficult to unravel; because everyone is susceptible, this leaves few points of difference to begin teasing apart the threads (Fischhoff, 1988). As a result, Gigerenzer (2000, p. 246), somewhat facetiously, has described overconfidence as a "fact waiting for a theory".

A common, starting assumption is that when overconfidence occurs, it indicates that the range was not wide enough to include the true value; although this has been questioned with Moore, Carter, and Yang (2015), for example, arguing from a statistical basis that the effect is primarily driven by errors in the location of estimates rather than range widths. Despite the lack of agreement, there are numerous theories relating to overconfidence; one theory suggests that overconfidence is a result of the way questions are asked and states that overconfidence disappears when confidence is asked for in terms of frequency, rather than probability (Gigerenzer, 2000; Gigerenzer, Hoffrage, & Kleinbolting, 1991). Another theory is that of the naïve intuitive statistician, which proposes that overconfidence results from people intuitively applying sample characteristics (estimated from the necessarily limited samples they can draw from memory) to populations (Juslin, Winman, & Hansson, 2007). However, other studies have failed to find these same effects - for both the probability-frequency and naïve intuitive statistician theories (e.g., Ferretti et al., 2016).

A complicating factor is that knowledge or expertise can affect overconfidence but not necessarily in a consistent manner; Block and Harper (1991), for example, state that people use narrower intervals, but are less overconfident estimating values they have greater knowledge of; Bruza, Welsh, Navarro, and Begg (2011) demonstrated that narrower ranges were associated with greater knowledge; and McKenzie, Liersch, and Yaniv (2008) showed that while experts were willing to provide narrower ranges, this resulted in them being just as overconfidence as less knowledgeable people. Moore et al. (2015) provide an overview of the various, sometimes contradictory research relating to overconfidence. This thesis deals with three potential causes of overconfidence which are addressed in the following sections.

1.3.1 Anchoring

Overconfidence was initially described by Tversky and Kahneman (1974) as an effect of the anchoring-and-adjustment heuristic. Specifically, when people are asked to provide a range, they start with their best estimate and adjust away from it to produce the end-points of their range, but they fail to adjust sufficiently, resulting in overconfidence. Anchoring, when considered in isolation of overconfidence, is a robust finding in decision making (Furnham & Boo, 2011); however, there is debate about anchoring as a cause of overconfidence. Block and Harper (1991) found that an anchor that was presented to participants correlated with their subsequent estimates, but did not increase overconfidence. Juslin, Wennerholm, and Olsson (1999) suggest that anchoring accounts for only a small amount of overconfidence.

1.3.2 Informativeness-Accuracy Trade-Off

The Informativeness-Accuracy Trade-Off (IAT; Yaniv & Foster, 1995) is based on conversational norms that suggest communication should be suitably informative and accurate (Grice, 1975). The width of intervals that are produced during elicitation tasks requires compromise between these two goals; while wider intervals are more likely to contain the true value, narrower intervals are more informative of what the person believes, due to the restricted content. This is the essence of the IAT and research has demonstrated that people are willing to sacrifice some amount of accuracy in order to increase informativeness (Yaniv & Foster, 1995). A factor that may contribute to the IAT is the timing of assessment of information; the utility of a narrow interval is immediately apparent, whereas the accuracy will most likely be assessed at a future date (Yaniv & Foster, 1997). An implication of the IAT is that receivers utilise the 'graininess' or width of intervals to evaluate the confidence with which the information is being presented; wider, grainier intervals indicate lower confidence in the value of interest (Welsh, Navarro, & Begg, 2011). The significance of IAT to overconfidence relies on the assumption that overconfidence is caused by overly narrow range widths (Moore & Healy, 2008).

1.3.3 Elicitation format

Format dependence, as described by Winman, Hansson, and Juslin (2004) refers to the impact that elicitation format has on overconfidence. A study was conducted that included the following tasks: in one condition, participants were asked to produce a range for a particular confidence level; and, in the second, participants were asked to evaluate a pre-stated interval. Production of intervals resulted in extreme overconfidence; however, this overconfidence was greatly reduced when participants evaluated pre-stated intervals instead (Winman et al., 2004). The effect of format dependence relates to Brunswikian notions that describe probability and frequency tasks as relying on different reference classes, which will provide different cues from one's natural environment to aid in assessing probability (Gigerenzer et al., 1991).

1.4 Individual Differences

Research into individual differences has the potential to identify stable characteristics that reduce susceptibility to biases (see, e.g., Schaefer, Williams, Goodie, & Campbell, 2004; Welsh & Begg, 2016; Welsh, Delfabbro, Burns, & Begg, 2014). There has been considerable investigation into individual differences and overconfidence, with some studies finding relationships, and others not (Moore & Healy, 2008). This thesis will investigate three individual difference measures: Need for Closure, Openness and Conscientiousness, which are described in the following sections.

1.4.1 Need for Closure

Need for Closure (NFC) is characterised by an urgent desire to reach a decision and once a decision has been made, to disregard any contradictory information (Roets & Van Hiel, 2007). NFC is also related to aversion to ambiguity and close mindedness (Webster & Kruglanski, 1994). The NFC short scale (which will be used in data collection herein) has also been found to correlate with openness ($r = -.47$, $p < .001$) and conscientiousness ($r = .31$, $p < .001$) (Roets & Van Hiel, 2011).

In terms of overconfidence, the aversion to ambiguity that characterises NFC has been related to interval production; given that wide intervals are more ambiguous than narrow ones, people high in NFC may produce narrower intervals. Kaesler, Welsh, and Semmler (2016) found evidence for such a correlation with higher NFC relating to narrower ranges ($r = 0.4$, $p < .05$). Assuming that overconfidence is the result of overly narrow ranges, NFC could, therefore predict susceptibility to overconfidence.

1.4.2 Openness

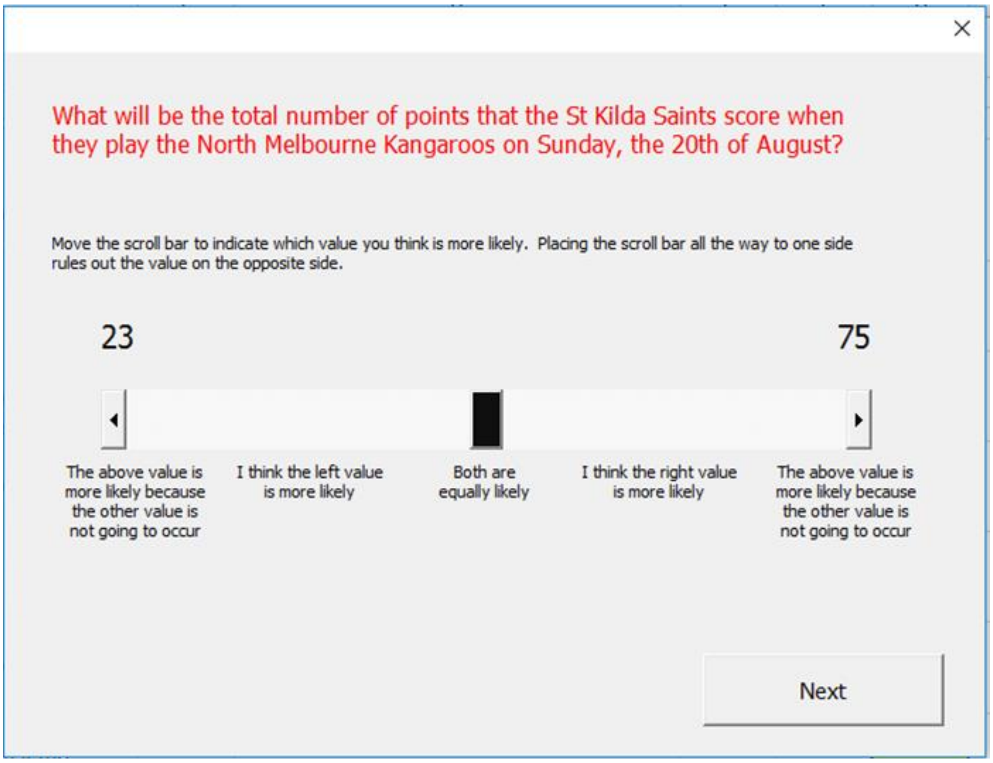
Openness seems to lend itself to decision making due to its very nature - describing an individual’s tendency to consider new information and to alter their own beliefs (McElroy & Dowd, 2007). This seems to align closely with anchoring, which describes the impact that received information has on a person (Tversky & Kahneman, 1974). In a study by McElroy and Dowd (2007), a significant relationship was found between anchoring and openness, with participants who were higher in openness being more affected by anchoring cues. However, a more recent study by Furnham, Boo, and McClelland (2012) did not find any relationship between anchoring and openness.

1.4.3 Conscientiousness

Conscientiousness is positively correlated to job and academic performance, and negatively correlated to intelligence; this apparently contradictory relationship is explained by people of lower intelligence compensating with conscientious behaviour, such as being thorough, persistent and methodical (Moutafi, Furnham, & Paltiel, 2004; Rammstedt, Danner, & Martin, 2016). People high in conscientiousness are also more likely to think carefully before making judgments (Furnham & Boo, 2011). Although conscientiousness has not been found to relate to overconfidence (Schaefer et al., 2004) or anchoring (Furnham et al., 2012), the characteristics of conscientiousness appear to be relevant to decision making.

1.5 What is the MOLE?

As noted above, the degree of overconfidence in a person's estimates can vary as a result of how questions are asked (Soll & Klayman, 2004; Winman et al., 2004). More or Less Elicitation (MOLE) is a computerised elicitation tool, which asks for estimates in a particular manner, with the aim of decreasing overconfidence (Welsh & Begg, 2016, Under review; Welsh, Lee, & Begg, 2008, 2009). The MOLE presents users with two random values selected from a pre-defined widest possible range (i.e., a range wide enough that it would contain any values that may occur). The user indicates which of these two values they believe to be closer to what the true value is (or will be). They do this by adjusting a scroll bar, where the extreme ends indicate 100% confidence in the associated value and the centre of the scroll bar indicates 50% confidence, that is, a belief that either value is equally likely. The user interface is shown in Figure 1. This process is repeated numerous times for each question with the number of iterations set in advance by the experimenter (ten being a commonly used value in previous studies). After each iteration, the program assesses the participant's response and may truncate the range. Specifically, if a user selects one of the values with absolute certainty, the program will then truncate the range by removing values that the user has indicated are not possible.



What will be the total number of points that the St Kilda Saints score when they play the North Melbourne Kangaroos on Sunday, the 20th of August?

Move the scroll bar to indicate which value you think is more likely. Placing the scroll bar all the way to one side rules out the value on the opposite side.

23
75
The above value is more likely because the other value is not going to occur
I think the left value is more likely
Both are equally likely
I think the right value is more likely
The above value is more likely because the other value is not going to occur
Next

Figure 1 MOLE user interface

At the end of the elicitation process, the MOLE program produces a final range containing all values that have not explicitly been ruled out and a best estimate calculated from the individual confidence judgements. The best estimate at each iteration is calculated by multiplying the confidence, converted to a 0 to 100 scale, by the difference between the two random values and adding to the lower value. That is, for two random values of 100 and 200 and a confidence of 50% (which indicates the user believes both values are equally likely), the best estimate will be 50% times the range (or difference between the two values) added to the lower value. In this case it results in a best estimate of 150. The best estimates from all iterations are averaged to produce a final best estimate.

1.5.1 Theoretical underpinnings

There are four key insights into decision making that are utilised by the MOLE. The first is to prevent anchoring. Anchoring has the potential to bias elicitation responses, by moving estimates closer to a number that a person has recently been exposed to. The MOLE guards against anchoring by providing numerous values, preventing a single value from having an undue effect. Additionally, the values presented by the MOLE – being randomly selected – may include ones that would not otherwise have been considered by the user. This strategy, of considering alternative or contradictory information has been shown to reduce anchoring (Mussweiler, Strack, & Pfeiffer, 2000; Russo & Schoemaker, 1992).

The second insight the MOLE uses is retaining uncertainty. The anchoring and adjustment theory proposes that people adjust from a starting value, but stop once the reach a value that they believe is appropriate, even though they would also consider subsequent values equally appropriate (Kahneman, 2011). Figure 2(a) illustrates how when starting from a best estimate in the middle of a range and working outwards, people stop once they approach the inner edges of their low or high value areas, resulting in a narrow range. The MOLE requires users start at the lower and upper bounds of possibility and then adjust inwards. Figure 2(b) shows how approaching their low or high values from the outside can, therefore, result in users stopping at the outside edges of their low or high ranges and, subsequently, producing wider ranges (Welsh & Begg, 2015).

The third insight utilised by the MOLE is that people are better at relative judgments than they are at absolute judgements. Soll and Klayman (2004) found that choosing between two options resulted in less overconfidence than interval estimates for similar questions. The MOLE utilises this ability to discern between values by presenting two options and having a person decide which is more likely before translating this information into absolute judgments.

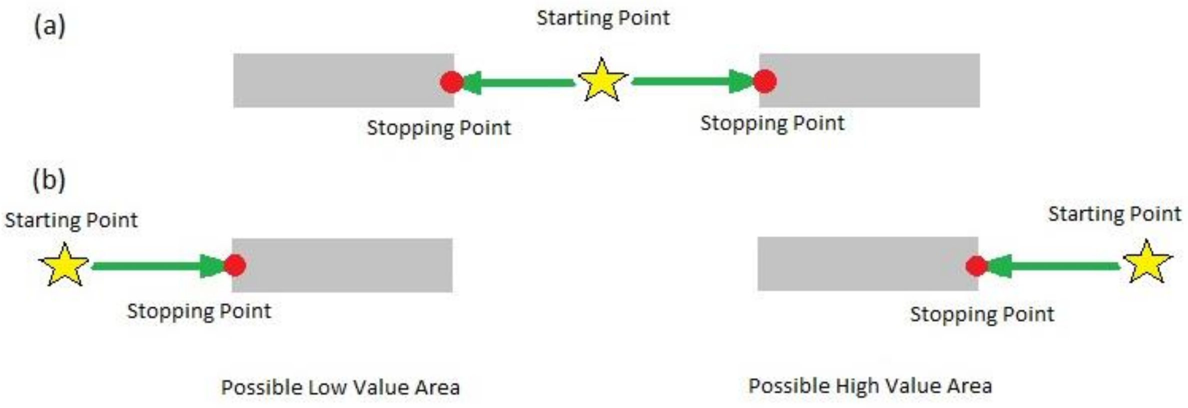


Figure 2 Stopping points in range production for (a) direct elicitation and (b) the MOLE, adapted from Welsh and Begg (2015)

The fourth and final insight that the MOLE uses is that the average of repeated estimates is more accurate than a single estimate. Herzog and Hertwig (2014) have demonstrated that asking a person a question numerous times and averaging the results provides a better estimate than a single answer, providing the estimates were, to some extent, independent. The MOLE's structure enables multiple, independent estimates to be made regarding the same parameter. All of these insights contribute to the design and efficacy of the MOLE program.

1.5.2 The benefits of the MOLE

Previous studies have demonstrated the benefits of the MOLE program in markedly reducing overconfidence and producing best estimates that correlate strongly with actual values; these effects have been demonstrated with perceptual, epistemic and forecasting tasks (Welsh & Begg, 2015, Under review; M. B. Welsh et al., 2008, 2009). The MOLE is also theorised to prevent anchoring on a specific value but this has not been definitively tested (Welsh & Begg, Under review).

1.6 Current Study

As described above, overconfidence is complex and poorly understood, with a number of areas of potential investigation. The current study includes a variety of factors potentially related to overconfidence, while focusing on the MOLE process. The MOLE will be investigated, with the intention of replicating previous findings and comparing performance of the MOLE to that of direct elicitation. While the MOLE has been assumed to overcome the effects of anchoring, this has not been tested; the current study will investigate this assumption. The study will also seek to understand the impact of format dependence on overconfidence when using the MOLE and direct elicitation. These investigations will be conducted in an experimental component of the study.

The study will also incorporate a correlational component, which will investigate the relationship of individual differences with performance on elicitation tasks. NFC has previously been associated with range width in a study that utilised a direct elicitation method, with participants being asked to provide a lower and upper estimate of an 80% confidence interval (Kaesler et al., 2016). It is anticipated that this effect could reverse in the presence of the MOLE due to the MOLE's requirement to rule out values. NFC's aversion to ambiguity in this case may be better served by retaining wide ranges.

By definition, openness seems relevant to anchoring, McElroy and Dowd (2007) found this to be the case, but other studies have failed to replicate the effect (Furnham et al., 2012). This study attempts to replicate the findings using the same measure, the Ten Item Personality Inventory (TIPI) and also includes a more robust measure, the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 2003).

Conscientiousness has not previously been found to relate to decision making (Furnham et al., 2012; Schaefer et al., 2004); however, conscientiousness has been shown to correlate to job and academic performance (Moutafi et al., 2004; Rammstedt et al., 2016). It is likely that the behavioural aspects of conscientiousness would be advantageous to the MOLE, due to it being a repetitive and lengthy task. It is anticipated that participants who are higher in conscientiousness will remain committed to the task, taking the care that is required to ensure optimal results.

1.6.1 Hypotheses

Hypothesis 1: Range widths produced by the MOLE will be significantly wider than those produced by direct elicitation.

Hypothesis 2: MOLE ranges will contain actual values more often than directly elicited ranges.

Hypothesis 3: MOLE best estimates will positively correlate with actual values.

Hypothesis 4: Anchor values presented prior to direct elicitation tasks will cause an anchoring effect - as demonstrated by best estimates correlating with anchor values.

Hypothesis 5: There will be no evidence of anchoring when anchor values are presented prior to MOLE tasks - as demonstrated by no correlation between best estimates and anchor values.

Hypothesis 6: There will be evidence of format dependence, with participants being better able to evaluate ranges than produce them.

Hypothesis 7: There will be a significant positive correlation between openness scores and anchoring effects.

Hypothesis 8: There will be a negative correlation between NFC scores and range width when direct elicitation is used.

Hypothesis 9: There will be a positive correlation between NFC scores and range width when the MOLE is used.

Hypothesis 10: Conscientiousness will increase the care participants take when completing the study, resulting in a positive correlation between conscientiousness and time taken to complete the survey.

Hypothesis 11: Conscientiousness will decrease careless responding, resulting in a negative correlation between conscientiousness and the number of 50% responses in MOLE conditions.

2 Method

2.1 Participants

The study included *N* = 62 participants (38 females and 24 males) aged between 18 and 65 years (*M* = 31.15, *SD* = 12.81). One participant was excluded from the study due to inappropriate responding. The participants included domestic students (*n* = 20), international students (*n* = 13) and people who were not currently studying (*n* = 23). The participants primarily reported English as their first language, *n* = 12 indicated that English was not their first language. Participation was restricted to people 18 years or older, able to read and write in English and currently living in Australia.

A statistical power analysis was conducted in G*Power, with α = .05 and power = .8. Previous MOLE studies reported large effect sizes(Welsh & Begg, Under review; M. B. Welsh et al., 2008). It was determined that 62 participants should be more than adequate to detect large effect sizes when conducting repeated measures ANOVA.

Participants were recruited within the University of Adelaide and from the wider population. Within the university, participants were recruited through flyers (see Appendix A) placed around the university and through the Psychology School's research participation website. The (*n* = 31) participants who responded to the flyers received a $20 gift card for completing the survey, while the (*n* = 14) first year psychology students received course credit for participation. Participants from outside the University *(n* = 17) were recruited through Facebook posts and a website and a $100 gift card was promised to the best performing participant.

2.2 Materials

The study consisted of various measures and elicitation tasks that were included in a single program, created in Visual Basic for Applications (VBA) in Microsoft Excel. The program was built specifically for the purposes of this study, integrating all aspects of the study into a single program, a copy of which is in Appendix B. Previous studies had used a version of the MOLE programmed in MATLAB, which required laboratory-based studies due to software requirements (Welsh & Begg, Under review). The benefit of the new program was that it could be administered remotely due to the portability and robustness of the program. The program was portable due to the ubiquity of Microsoft programs and was robust, because it prevented users from skipping fields and restricted the input to ensure that questions were answered appropriately. Allowing participants to complete the survey remotely was anticipated to increase recruitment numbers.

2.2.1 Measures

The survey collected demographic information; measures of need for closure, openness and conscientiousness; and five elicitation tasks. Details of the measures are included in the following sections.

2.2.1.1 Demographic

Participants were asked to provide information about themselves, including age, gender and their engagement with Australian Rules Football (ARF). Engagement with ARF was measured using four questions, such as how often participants played ARF or watched ARF matches. Participants rated their frequency of engagement for each of four questions on a four-point Likert scale, with a response of one corresponding to '*rarely or never*' and four corresponding to '*more than once a week'*. Possible scores for engagement with ARF ranged between 4 and 16 with higher scores indicating greater engagement.

2.2.1.2 Need for closure

Participants completed the 15 item Need for Closure (NFC) Scale (Roets & Van Hiel, 2011). NFC items include "I don't like situations that are uncertain" and "When I have made a decision, I feel relieved". Participants rated the extent to which they agreed with each of the 15 items on a six-point Likert scale, where one corresponded to a response of '*completely disagree*' and six corresponded to a response of '*completely agree*'. Possible scores thus range between 15 and 90, with higher scores indicating higher NFC. The 15 item NFC scale has been shown to have appropriate reliability, with Cronbach's alpha = .87 and test-retest reliability of *r* = .79 (Roets & Van Hiel, 2011).

2.2.1.3 Ten Item Personality Inventory - openness subscale

The openness subscale of the Ten Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003) was included in the survey. Participants were asked to rate the degree to which they agreed with two statements on a Likert scale from one (*disagree strongly*) to seven (*agree strongly*). The TIPI directly asks about personality traits, rather than about facets of personality, specifically asking participants to rate how 'open' they are (Gosling et al., 2003). Possible scores range between two and 14, with higher scores indicating greater openness. While Cronbach's alpha is not a suitable measure due to the presence of only two items in the scale, the TIPI openness subscale has a test-retest reliability of *r*= .62, and correlates at *r* = .65 with the Big-Five Inventory openness subscale (Goldberg, 1992; Gosling et al., 2003).

2.2.1.4 NEO-FFI openness and conscientiousness subscales

The survey included the openness and conscientiousness subscales of the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 2003). The openness subscale asks participants to consider statements such as "I have a lot of intellectual curiosity" and "I often try new and foreign foods", while the conscientiousness subscale includes statements such as "I keep my belongings neat and clean" and "I work hard to accomplish my goals". Each subscale included 12 statements with participants asked to rate the extent to which they agreed with each statement on a five-point Likert scale. The scale ranged from '*strongly* *disagree*', with a score of zero, to '*strongly agree*', with a score of 4. Total possible scores for both subscales thus vary between 0 and 48. Higher scores indicate greater openness in the openness subscale and greater conscientiousness in the conscientiousness subscale. The Cronbach's alphas of the measures are .76 for openness and .84 for conscientiousness; the test retest reliability is .88 for openness and .90 for conscientiousness (Costa & McCrae, 2008)

2.2.2 Elicitation tasks

The survey included five different elicitation tasks. The tasks were (1) More Or Less Elicitation (MOLE), (2) MOLE with anchoring, (3) direct elicitation, (4) direct elicitation with anchoring and (5) a summary task. The study utilised forecasting questions, specifically related to Australian Football League (AFL) match results. A forecasting approach was chosen due to the study being conducted remotely; epistemic questions can be easily answered with an internet search, whereas forecasting questions relate to genuinely unknowable values. AFL results were chosen as the forecasting measure because they provided a sufficient amount of numerical values of similar difficulty to predict. Participants were asked to consider the total number of points that particular teams would score in the next two rounds. For example: "What will be the total number of points that the St Kilda Saints score when they play the North Melbourne Kangaroos on Sunday, the 20th of August?". Given that the study was conducted in Adelaide, the two Adelaide-based AFL teams were excluded in an attempt to limit the impact of specialist knowledge.

2.2.2.1 MOLE

The MOLE task utilises the MOLE process (Welsh & Begg, Under review). Five questions, with ten iterations of each, were included in the MOLE elicitation task. At each iteration of the MOLE process, participants were presented with two values that were randomly selected from a pre-determined widest possible range (in the case of this study, 0 to

300 points). Participants were asked to indicate which value they believed would be closer to the true value by adjusting a scroll bar from the default centre position, which indicates that the participant believes that both of the values presented are equally likely to occur (user interface shown in Figure 1). Adjusting the scroll bar to the extreme left indicated that the value displayed on the right was not possible and vice versa, whereas positions intermediate between the centre and the end-points were mapped onto levels of confidence between 50% and 100% in the selected value being closer to the true value

When a participant indicated maximum (100%) confidence in one of the values, the possible range was truncated at the unselected value. For example, given an initial range of 0 to 300, if the values 95 and 240 were presented and the participant selected 95 with maximum confidence, then this would result in the range being truncated at 240 and the options displayed in later iterations being drawn from the 0-240 range rather than 0-300. If confidence was less than maximum, the range remained unchanged. (NB - previous use of the MOLE truncated the range at the midpoint, rather than the unselected value (Welsh & Begg, Under review). This study adopted a more conservative approach to range truncation due to overly narrow ranges in a previous study.)

The range containing all the values that remained after ten iterations of the MOLE was the 100% confidence range. Participants are assumed to be 100% confident that this will contain the true value, having not ruled any of the values out. Participant's best estimates of the true value were calculated at each iteration by multiplying the confidence by the difference between the values as explained in section 1.5. A participant's overall best estimate was simply the average of their best estimates at each iteration.

2.2.2.2 MOLE with anchoring

The MOLE with anchoring task used the same format as the MOLE task, with the addition of an anchoring question presented prior to each of five different MOLE questions (again, each consisting of 10 iterations). The anchoring question asked if the team in question would score more/less than a series of values that were linearly distributed over the initial range (i.e.: 20, 90, 160, 230 or 300). An example of the anchoring question is “Will the Collingwood Magpies score less than 160 total points when they play the Geelong Cats on Saturday, the 19th of August?”. In this example, the value of 160 is intended to be an anchor. Participants completed five MOLE with anchoring tasks for which range widths and best estimates were calculated in the same manner as for the MOLE task.

2.2.2.3 Direct elicitation

The direct elicitation task required participants to enter a minimum value, a maximum value and their best estimate of the points that the indicated team would score. The program checked that the minimum value was lower than the best estimate, which in turn had to be lower than the maximum; failure to meet these requirements resulted in the program displaying an error message.

2.2.2.4 Direct elicitation with anchoring

The direct elicitation task with anchoring was the same as the direct elicitation task, but it also first asked if the team in question would score more/less than one of a series of anchoring values from within the widest possible range used by the MOLE.

2.2.2.5 Summary elicitation task

The final elicitation task asked participants to evaluate the ranges that they, themselves had produced in the first four elicitation tasks. The ranges were presented to participants, although they were not explicitly identified as the ranges they had produced.

Participants were then required to rate how confident they were, from 0% to 100%, that the true value would lie within their stated range, for each of the 20 elicited ranges across the four different tasks.

2.3 Procedure

After participants had registered their interest in the study, they were emailed a copy of the VBA program. Participants were provided with a program that included questions about the next two (weekly) rounds of AFL matches. Therefore, the specific questions included in the program changed on a weekly basis. Results were collected between the 5th of June 2017 and the 19th of August 2017. The study utilised a within-participants design, with all participants completing all elicitation tasks, allowing for comparison of performance across the tasks.

The survey began with an information page, where participants clicked a button to indicate their agreement to participate and to begin the survey. The survey progressed through, in order, demographic questions, NFC, TIPI openness subscale, NEO-FFI openness subscale, and the NEO-FFI conscientiousness subscale prior to beginning the four elicitation tasks and then concluded with the summary elicitation task and instructions to save the file and return it to the researcher by email. Participants could only move forwards through the tasks, once a response was submitted it could not be altered. When the completed surveys were received by the researcher, they were saved with an identification code to ensure that results were anonymous.

Of the elicitation tasks (i.e. MOLE, MOLE with anchoring, direct elicitation, direct elicitation with anchoring and the summary task), the summary task was necessarily always performed last while a balanced Latin square was used to determine four different orders for the remaining elicitation tasks as shown in Table 1. Participants were randomly assigned to one of these four conditions, with the intention of preventing order effects. The number of completed surveys in each condition for each date is shown in Table 2. While efforts were made to allocate participants equally to the four conditions, the number of participants varied due to uptake and completion rates.

Table 1 Order of tasks by condition

	Condition 1	**Condition 2**	**Condition 3**	**Condition 4**
1st	MOLE	Direct with anchor	MOLE with anchor	Direct
2nd	Direct	MOLE with anchor	MOLE	Direct with anchor
3rd	MOLE with anchor	Direct	Direct with anchor	MOLE
4th	Direct with anchor	MOLE	Direct	MOLE with anchor

Table 2 Survey completion rates

	Closing date of survey										
	June			**Jul**				**Aug**			
Condition	**7**	**15**	**30**	**8**	**14**	**22**	**28**	**4**	**11**	**19**	**Total**
1				2	3	1	4	4	2	5	21
2	1	1	1	1			1	4	2	3	14
3		2	1	1	1		2	1	1	3	12
4		1	2	2				4	3	3	15
Total	1	4	4	6	4	1	7	13	8	14	62

3 Results

This study employed a two-by-two, within participants design, which compared two types of elicitation (MOLE and direct elicitation) in the presence or absence of anchors. Table 3 shows the resulting four conditions and the abbreviations used to describe them. All analysis was performed in R, with two-tailed tests, unless specified.

Table 3 Study conditions

	Presence of Anchor	
Elicitation Method	**Anchor**	**No Anchor**
MOLE	AMOLE	NMOLE
Direct Elicitation	ADIR	NDIR

3.1 Dependent Measures

The dependent measures of the study are defined as follows:

- Range width: the distance between the minimum and maximum elicited values.
- Calibration: the percentage of an individual's ranges that contain the actual value of interest.
- Best estimate: the value that the participant indicates is most likely to represent the actual value.
- Error: the average distance between best estimates and the actual value, calculated for each participant.
- Accuracy score: the correlation between best estimates and actual values across the five questions within each condition for each participant.

3.2 Descriptive Statistics

A summary of descriptive statistics is provided in Table 4. This includes the predictor variables (demographic and individual differences) and the dependent measures for each condition. As the study investigated how accurately participants could forecast the final score in AFL matches, the descriptives for the set of matches used during the study are

Table 4 Descriptive Statistics

Continuous Variables	**Mean**	**min**	**max**	**SD**
Time to complete	00:30:18	00:08:45	01:24:26	00:16:59
Age	31.15	18	65	12.81
Engagement with ARF	7.61	4	16	2.85
Need for Closure	57.69	41	80	9.12
TIPI O	9.69	5	13	2.16
NEO-FFI O	29.82	18	43	5.95
NEO-FFI C	30.11	10	44	6.85
Range width				
ADIR	76.42	13	330	55.41
AMOLE	204.96	55.4	300	61.73
NDIR	73.58	17.8	320	51.47
NMOLE	203.99	87.4	300	63.42
Calibration				
ADIR	57%	0%	100%	31%
AMOLE	93%	40%	100%	14%
NDIR	60%	0%	100%	26%
NMOLE	94%	60%	100%	12%
Post hoc calibration				
ADIR	67%	17%	99%	17%
AMOLE	84%	1%	100%	22%
NDIR	67%	16%	100%	16%
NMOLE	83%	15%	100%	20%
Accuracy score				
ADIR	.02	-.92	.94	.54
AMOLE	.17	-.96	.94	.50
NDIR	.21	-.70	.99	.54
NMOLE	-.01	-.90	.90	.49
Anchoring score				
ADIR	.14	-.96	.99	.50
AMOLE	.01	-.96	.93	.51
Average error				
ADIR	39.87	6.80	226.4	36.00
AMOLE	46.32	8.13	120.39	26.63
NDIR	35.26	9.40	162.80	26.35
NMOLE	43.50	11.40	111.40	23.04
Average best estimate				
ADIR	106.83	59.20	330	43.48
AMOLE	123.98	72.30	197.99	28.70
NDIR	99.91	56.8	264	37.38
NMOLE	122.70	81.8	196.40	26.64

included, here, for comparison: *M* = 87.70, *SD* = 23.59, min = 39, max = 163. Australian Tertiary Admission Rank (ATAR) or Special Tertiary Admissions Test (STAT) results were collected as an indicator of general intelligence, however only 20 of 62 participants provided appropriate responses; therefore, analysis was not conducted for this variable.

3.3 Preliminary Analysis

As a preliminary analysis, all variables were included in a correlation matrix. Given the large size of this matrix, it has been included in Appendix C as Table C1. Relevant sections of the matrix have been included in the following Results section.

3.4 Elicitation Effects

3.4.1 Range widths

Hypothesis one proposed that ranges produced using the MOLE would be significantly wider than those produced by direct elicitation. Figure 3 displays the range widths for each of the four conditions; inspection of Figure 3 suggests that the range widths are consistently wider for NMOLE and AMOLE conditions than the ADIR and DIR conditions. The plot also suggests the presence of outliers; however, they did not alter the significance of the results and were not removed. A two-way repeated measures ANOVA was conducted. Results of the analysis confirmed that there was a significant difference, with a very large effect size, between the widths of ranges elicited in the MOLE conditions and those elicited in the direct conditions $F(1,61) = 235$, $p < .001$, $\eta_p^2 = .79$. The presence of an anchor did not result in a significant difference, nor did the interaction between elicitation type and presence of anchor. These results are consistent with the first hypothesis that the MOLE would result in wider ranges than direct elicitation. Further, it was found that range width was positively correlated to calibration with a medium effect size for all conditions. The results of these correlations for the various conditions are found in the following tables:

ADIR in Table 7, AMOLE in Table 8, NDIR in Table 9 and NMOLE in Table 10. This demonstrates that wider ranges contain the true value more often.

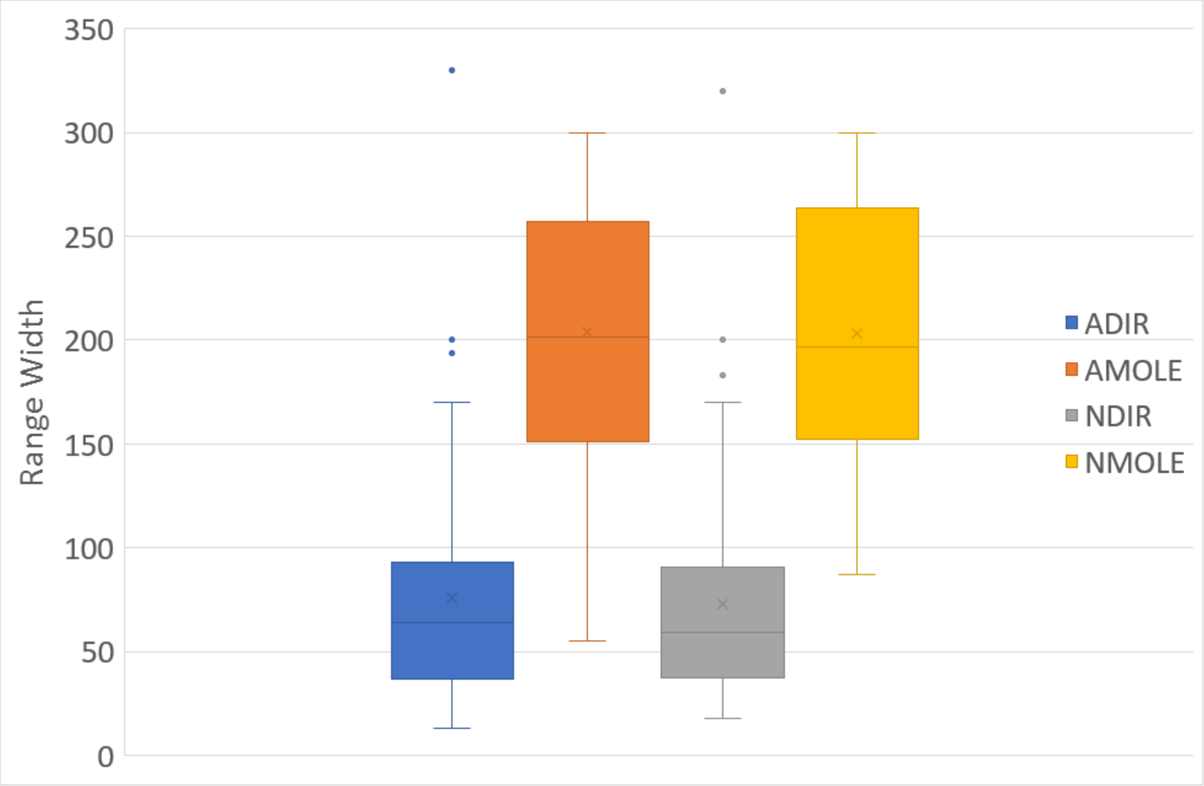


Figure 3 Box and whisker plot for range widths

3.4.2 Calibration

Hypothesis two proposed that calibration would be better for the MOLE than for direct elicitation. Calibration is expressed as the percentage of ranges that encompass the true result. The box and whisker plot in Figure 4 suggests that calibration is better for the NMOLE and AMOLE conditions. A two-way repeated measure ANOVA was conducted on calibration scores. There was a significant difference with a very large effect size in the calibration based on the elicitation method, $F(1,61) = 126$, $p < .001$, $\eta_p^2 = .67$. There was no significant difference based on anchoring or the interaction of anchoring and elicitation method. These results are consistent with our second hypothesis that calibration would be higher for the MOLE.

3.4.3 MOLE correlation to actual values

The third hypothesis was that the best estimates produced by the NMOLE and AMOLE conditions would be positively correlated to the actual values. A repeated measures correlation (r_{rm}) was used due to there being five data points for each participant. Repeated measures correlation is equivalent to Pearson's correlation; however, it accommodates multiple data points per participant (increasing power), but does not violate the assumption of independence (Bakdash & Marusich, 2017). Repeated measures correlation evaluates overall intra-individual relationships; it is calculated using a form of ANCOVA and can be performed in the rmcorr package in R (Bakdash & Marusich, 2017). Analysis showed a correlation between best estimates and actual values of $r_{\text{rm}}(247) = .14$, $p = .03$ for the AMOLE condition and $r_{\text{rm}}(247) = .001$, $p = .98$ for the NMOLE condition. The correlation for AMOLE was small and there was no correlation for NMOLE, when considered together, these results do not provide convincing evidence of a correlation between best estimates and actual vales. Therefore, the third hypothesis was not supported. Additionally, there was a small yet significant correlation between NDIR best estimates and actual values ($r_{\text{rm}}(247) = .16$, $p = .009$). There was not a significant correlation for the ADIR condition ($r_{\text{rm}}(247) = -.10$, $p = .13$).

3.5 Anchoring

Hypothesis four proposed that there would be a positive correlation between anchor values and the best estimate that participants produced in direct elicitation. Repeated measures correlation was used. Analysis confirmed a small to medium correlation of $r_{\text{rm}}(247) = .27$, $p < .001$. This suggests that anchor values relate to the best estimate, with higher anchors resulting in higher estimates.

Hypothesis five proposed that anchoring would not be evident in the AMOLE task. Repeated measures correlation analysis confirmed this with no correlation detected between AMOLE best estimates and anchor values, $r_{\mathrm{rm}}(247) = .02$, $p = .76$.

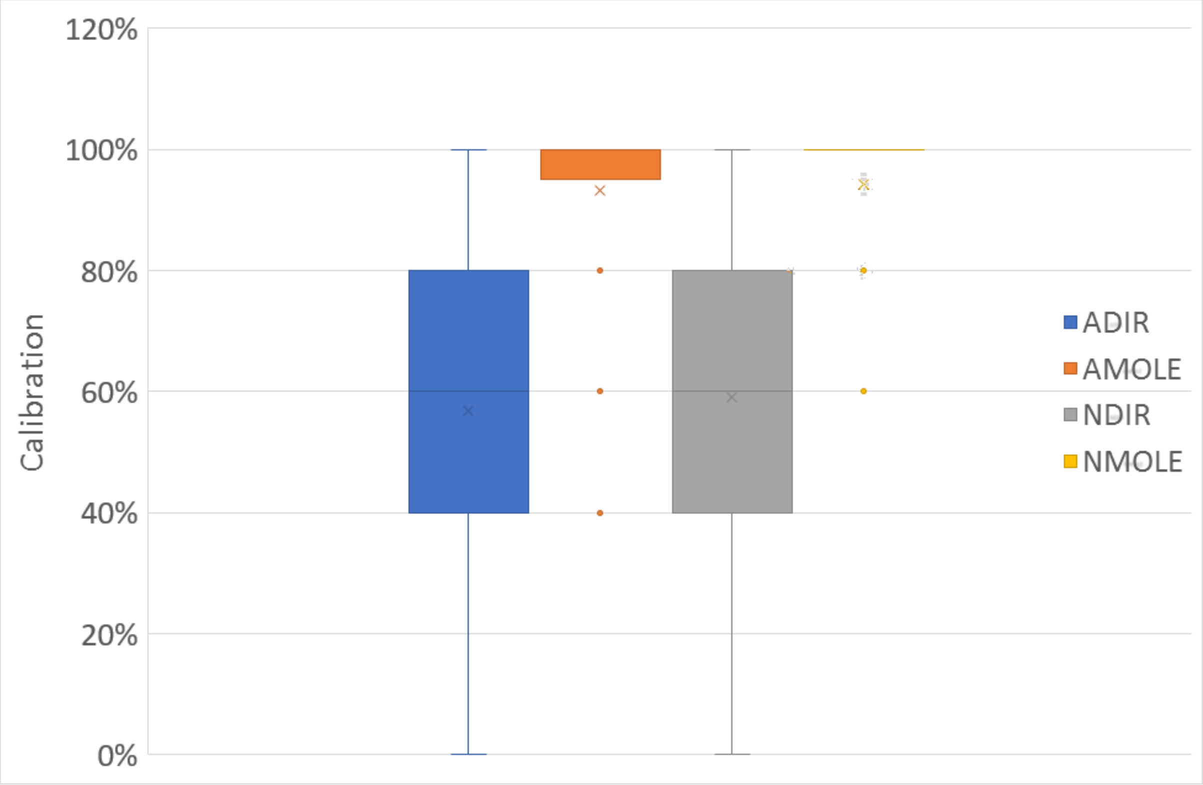


Figure 4 Box and whisker plots for calibration results

3.6 Format Dependence

Hypothesis six proposed that participants would be better able to evaluate ranges than to produce ranges. All the conditions elicited 100% confidence ranges; that is, the range within which the participant was 100% confident the true value would lie. After this, participants rated their confidence in the ranges that they had produced. Combining this information with the actual calibration scores provides two useful values: one is the difference between the expected 100% confidence and the actual calibration scores, the production error; the other is the difference between the actual scores and post hoc confidence levels, the evaluation error. These values were compared for each condition using a paired sample t-test, the results are shown in Table 5. For both the ADIR and NDIR conditions, the production error was significantly different than the evaluation error, with a very large effect size. Inspection of the means tells us that participants did not produce ranges that contained the true value as often as expected, but that they were better able to evaluate their own ranges to determine if they would contain the true value. The negative sign of the mean evaluation error indicates that the post hoc confidence levels were higher than the actual calibration scores (indicating overconfidence). This result was not repeated for the AMOLE and NMOLE conditions; there was no significant difference between the production error and the evaluation error. This indicates that participants were equally effective in producing ranges and evaluating those ranges for the AMOLE and NMOLE conditions. While the difference was not significant, the positive mean evaluation errors indicate that the post hoc confidence levels were lower than the actual calibration scores (indicating underconfidence). There were some unusual results, with some participants giving low confidence ratings to very wide ranges. These results did not impact the analysis and were included in the analysis.

3.7 Individual Differences

3.7.1 Openness and anchoring

Hypothesis seven proposed that higher openness scores would relate to greater anchoring effects. Openness was measured using both the Ten Item Personality Inventory (TIPI) openness subscale and the NEO Five-Factory Inventory (NEO-FFI) openness subscale. These two measures of openness were positively correlated, $r(60) = 0.49$, $p <.001$, as shown in Table 6. The effect of anchoring was quantified by an individual correlation score for each participant. This anchoring score was calculated as the correlation between anchor values and best estimates over the five questions in each of the anchor conditions.

There was no significant correlation between either openness measure and the anchoring score for the ADIR condition, shown in Table 7; nor was either openness measure significantly related to the anchoring score for the AMOLE condition, shown in Table 8. One-tailed correlations were used. Hypothesis seven was, thus, not supported as there was no evidence of increased anchoring for higher openness scores.

Table 5 Paired sample t-test results investigating format dependence

	Production error	**Evaluation error**				
Measure	**M%(SD)**	**M%(SD)**	***t* (df)**	**95% CI**	***p***	***d***
ADIR	0.43 (0.31)	-0.09 (0.29)	7.27 (61)	0.38 to 0.67	<.001	0.92
AMOLE	0.07 (0.14)	0.10 (0.25)	-0.68 (61)	-0.11 to 0.06	.50	0.09
NDIR	0.41 (0.27)	-0.07 (0.24)	7.93 (61)	0.36 to 0.61	<.001	1.01
NMOLE	0.06 (0.12)	0.12 (0.22)	-1.64 (61)	-0.13 to 0.01	.11	0.21

3.7.2 Need for closure and range width

Hypothesis eight proposed that Need for Closure (NFC) scores would be negatively correlated to range widths in the direct elicitation measures and hypothesis nine proposed that NFC would be positively correlated to range widths for MOLE measures. Figure 5 shows a scatterplot of these two variables for each condition with linear trendlines. Neither of the direct conditions had significant correlations between NFC and range width; ADIR (*r*(60) = .001, *p* = .52, one-tailed), NDIR (*r*(60) = -.03, *p* = .40, one-tailed). Therefore, hypothesis eight is not supported, with no correlation detected between NFC and range width for direct elicitation.

As suggested by Figure 5, there are positive relationships between NFC and range width for AMOLE (*r*(60) = .21, *p* = .048, one-tailed) and between NFC and range width for NMOLE (*r*(60) = .24, *p* = .03, one-tailed). These correlations indicate a small effect size and if they had been analysed with a two-tailed correlation, they would not be significant.

Therefore, there is tentative evidence to support hypothesis nine, that NFC relates to wider ranges when using the MOLE.

Inspection of Figure 5 also shows that all values of NFC are greater than 40, but it should be noted that the possible range is 15 to 90; therefore, the participants in this study only represent a subset of the possible scores for NFC. Results for the various conditions discussed in this section are found in the following tables: ADIR in Table 7; AMOLE in Table 8; NDIR in Table 9; and NMOLE in Table 10.

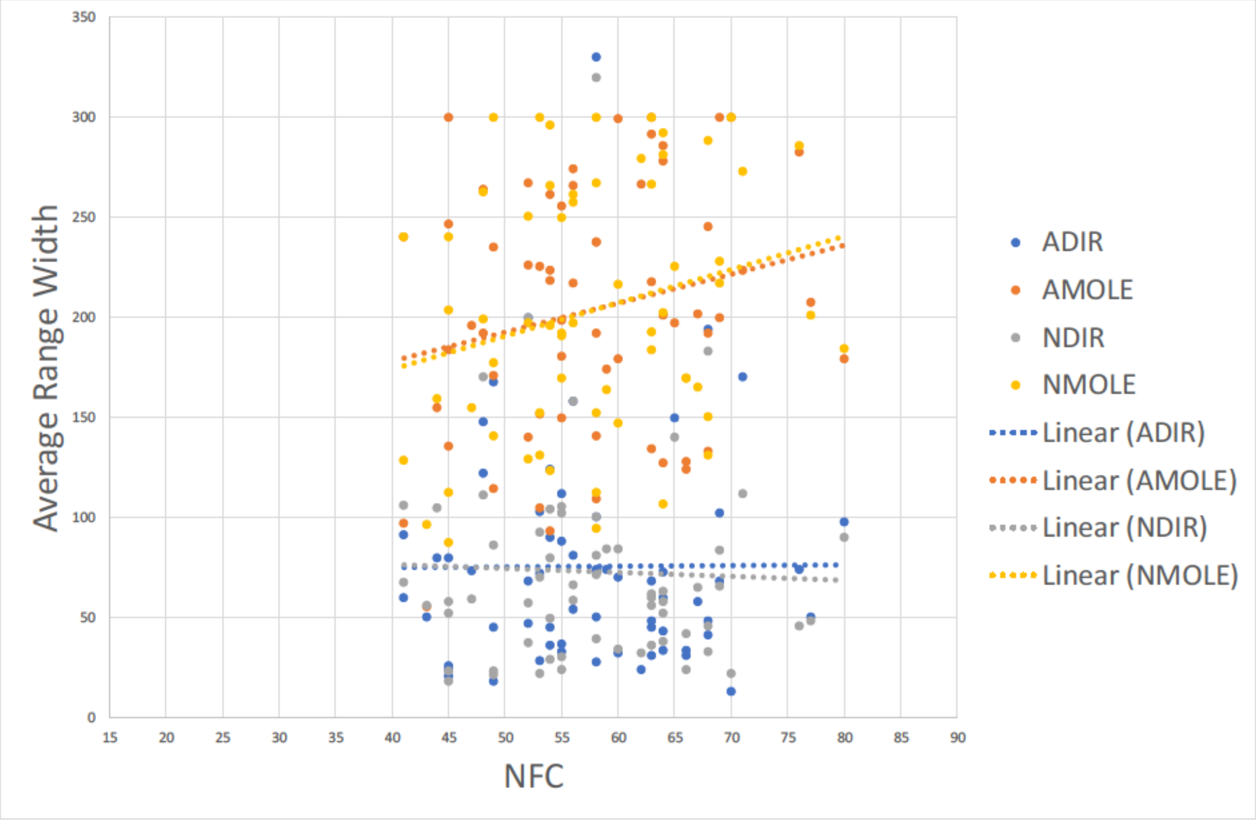


Figure 5 Scatterplot of NFC v average range width

3.7.3 Conscientiousness

Hypotheses ten proposed that conscientiousness would increase the care with which people approached the survey task, resulting in longer overall time to complete the survey. Analysis, however, found no correlation between conscientiousness and completion time ($r(60) = .01$, $p = .46$, one-tailed); therefore, Hypothesis ten was not supported. Additionally, Hypothesis eleven proposed that participants higher in conscientiousness would have fewer 50% responses to NMOLE and AMOLE tasks. The 50% response is the default and would, thus, occur more frequently when participants click through the responses without attempting to provide answers. Analysis showed that there was not a negative correlation (*r*(60) = .28, *p* = .99, one-tailed). This suggests the opposite of our hypothesis, participants higher in conscientiousness were more likely to make 50% responses; therefore, hypothesis eleven was not supported.

Table 6 Correlations of demographic and individual difference measures

Measures	**2**	**3**	**4**	**5**	**6**
1. Age	0.15	-0.06	0.04	0.05	0.10
2. Engagement with ARF	-	-0.27	0.17	-0.04	0.14
3. NFC		-	-0.45***	-0.14	0.16
4. TIPI_O			-	0.49***	-0.13
5. NEO-FFI_O				-	-0.11
6. NEO-FFI_C					-

Note: *p* < .05 = *, *p* <.01 = **, *p* < .001 = ***

Table 7 Correlations for ADIR condition

Measures	**7**	**8**	**9**	**10**	**11**
1. Age	0.42***	-0.20	-0.16	0.11	-0.30*
2. Engage ARF	0.44***	-0.46***	-0.16	0.35**	-0.41***
3. NFC	-0.14	-0.12	0.01	0.10	0.05
4. TIPI_O	0.29*	-0.03	-0.04	-0.07	-0.13
5. NEO-FFI_O	0.11	-0.07	-0.04	0.08	-0.16
6. NEO-FFI_C	0.07	-0.30*	-0.15	0.10	-0.20
ADIR conditon					
7. Accuracy score	-	-0.33**	-0.12	0.25*	-0.34**
8. Anchoring score		-	0.33**	-0.24	0.44***
9. Average range width			-	0.29*	0.64***
10. Calibration				-	-0.42***
11. Average error					-

Note: p < .05 = *, *p* <.01 = **, *p* < .001 = ***

Table 8 Correlations for AMOLE condition

Measures	**7.**	**8.**	**9.**	**10.**	**11.**
1. Age	-0.07	-0.05	-0.17	0.08	-0.22
2. Engage ARF	0.14	-0.15	-0.53***	-0.04	-0.60***
3. NFC	0.00	0.05	0.21[1]	0.16	0.13
4. TIPI_O	0.20	-0.16	-0.05	-0.14	0.04
5. NEO-FFI_O	0.10	-0.12	0.02	-0.17	0.06
6. NEO-FFI_C	0.10	-0.01	-0.03	-0.24	-0.03
AMOLE conditon					
7. Accuracy score	-	-0.13	-0.07	0.08	-0.24
8. Anchoring score		-	0.06	0.00	0.00
9. Average range width			-	0.44***	0.67***
10. Calibration				-	-0.17
11. Average error					-

Note: p < .05 = *, *p* <.01 = **, *p* < .001 = ***, *p* < .05 one-tailed = [1]

Table 9 Correlations for NDIR condition

Measures	**7.**	**8.**	**9.**	**10.**
1. Age	-0.02	-0.10	0.20	-0.24
2. Engage ARF	0.40**	-0.13	0.40**	-0.43***
3. NFC	0.06	-0.03	0.01	0.01
4. TIPI_O	0.11	-0.05	-0.07	-0.07
5. NEO-FFI_O	-0.03	-0.04	-0.01	-0.05
6. NEO-FFI_C	0.01	-0.20	0.01	-0.12
NDIR condition				
7. Accuracy score	-	-0.12	0.21	-0.36**
8. Average range width		-	0.33**	0.68***
9. Calibration			-	-0.30*
10. Average error				-

Note: *p* < .05 = *, *p* <.01 = **, *p* < .001 = ***

Table 10 Correlations for NMOLE condition

Measures	**7.**	**8.**	**9.**	**10.**
1. Age	0.04	-0.21	-0.07	-0.31*
2. Engage ARF	-0.20	-0.58***	-0.03	-0.47***
3. NFC	0.03	0.24[1]	0.12	0.19
4. TIPI_O	-0.08	-0.09	-0.15	0.01
5. NEO-FFI_O	-0.13	-0.04	0.03	0.07
6. NEO-FFI_C	0.14	-0.09	-0.24	0.09
NMOLE condition				
7. Accuracy score	-	0.17	-0.05	-0.04
8. Average range width		-	0.37**	0.61***
9. Calibration			-	-0.10
10. Average error				-

Note: *p* < .05 = *, *p* <.01 = **, *p* < .001 = ***, *p* < .05 one-tailed = [1]

3.8 Other Findings

3.8.1 Engagement with Australian Rules Football

Level of engagement with Australian Rules Football (ARF) was used to quantify participants knowledge about ARF, with higher engagement scores presumably reflecting greater knowledge. Engagement was significantly correlated with many other variables. Specifically, participants with greater ARF engagement performed appreciably better in many regards.

Results for the ADIR condition included several significant correlations with engagement, as shown in Table 7. For the ADIR condition, greater engagement related to higher accuracy scores ($r(60) = .44$, $p < .001$). The ADIR condition also demonstrated a negative correlation between ARF engagement and anchoring scores, ($r(60) = -.46$, $p < .001$). These results indicate that in the ADIR condition, people with higher knowledge of ARF were better able to predict the results of matches and were less susceptible to anchoring effects, both with medium effect sizes. Engagement was also positively correlated to calibration ($r(60) = .35$, $p < .01$), with a medium effect size. Given calibration refers to the percentage of ranges that contained the actual value, increased engagement was related to the production of ranges that were more likely to contain the actual value. Additionally, there was a negative correlation, ($r(60) = -.41$, $p < .001$) with a medium effect size between engagement and average error, the average distance between best estimates and actual values. That is, participants with greater engagement produced best estimates that were closer to the true value.

The AMOLE condition had two significant correlations with engagement, shown in Table 8. Engagement was negatively correlated to range width, ($r(60) = -.53$, $p < .001$) with a large effect size, indicating that participants with greater engagement reduced their ranges more. Engagement was also negatively correlated with average error, (*r*(60) = -.60, *p* < .001) with a large effect size.

The NDIR condition had 3 significant correlations involving engagement with ARF, shown in Table 9. The accuracy score was positively correlated to engagement, (*r*(60) = .40, *p* < .01) with a medium effect size, indicating that participants with higher engagement produced best estimates that were closer to the actual value. Engagement was also positively correlated to calibration, (*r*(60) = .40, *p* < .01) with a medium effect size, meaning that higher engagement was associated with more ranges that contained the actual value. Finally, for the NDIR condition, engagement was negatively correlated to average error, (*r*(60) = -.43, *p* < .001) with a medium effect size.

The NMOLE condition had two significant correlations involving engagement, shown in Table 10. Engagement was negatively correlated to range width, (*r*(60) = -.58, *p* < .001) with a large effect size, suggesting that participants with higher engagement produced narrower ranges. Engagement was also negatively correlated to average error, (*r*(60) = -.47, *p* < .001) with a medium effect size.

Overall, increased engagement with ARF had positive outcomes across all four conditions. In particular, a correlation between engagement and average error was present in all conditions, indicating that participants higher in engagement consistently produced best estimates that were closer to the actual value.

3.8.2 Openness

There was a significant relationship between TIPI openness and accuracy scores for the ADIR condition (*r*(60) = .29, *p* = .02, two-tailed). A significant relationship was not found in any other conditions, nor was there a relationship with the other measure of openness, the NEO-FFI; this suggests the possibility of a type I error. Results are shown in Table 7.

3.8.3 Conscientiousness

In the ADIR condition there was a negative correlation between conscientiousness and anchoring ($r(60) = -.30$, $p = .02$, two-tailed). This suggests that the effect of anchoring was less pronounced in participants with higher conscientiousness scores.

3.8.4 Error

An average error score was calculated for each participant under each of the four conditions, a summary of which is shown in Figure 6. Inspection of the chart suggests the presence of outliers, which were found to affect analysis. Therefore, three participants were removed from the data set due to producing values more than three times the interquartile range above the 75th percentile (Tukey, 1977). A two-way repeated measure ANOVA was conducted in R and found significant difference between direct and MOLE elicitation methods ($F(1,58) = 18.42$, $p < .001$, $\eta_p^2 = .24$) and between anchor and no anchor conditions ($F(1,58) = 4.69$, $p = .03$, $\eta_p^2 = .07$). A post hoc pairwise t-test with Holm correction determined that there was a significant difference in the error between the NDIR and NMOLE conditions ($p < .001$), NDIR and AMOLE ($p < .001$) and ADIR and AMOLE ($p = .02$).

Within each condition, average error was significantly correlated to range width, as shown in Tables 5 to 8. This result was consistent across the conditions, with correlations varying between $r = .61$ and $r = .68$. This is illustrated in Figure 7, with all of the trendlines having visibly similar slopes. The scatterplot also demonstrates that while the relationship is consistent, the effect is different for the different conditions. In section 3.4.1, it was established that the MOLE conditions had significantly wider ranges; therefore, the effect of increasing error as range width increases is amplified in the MOLE conditions.

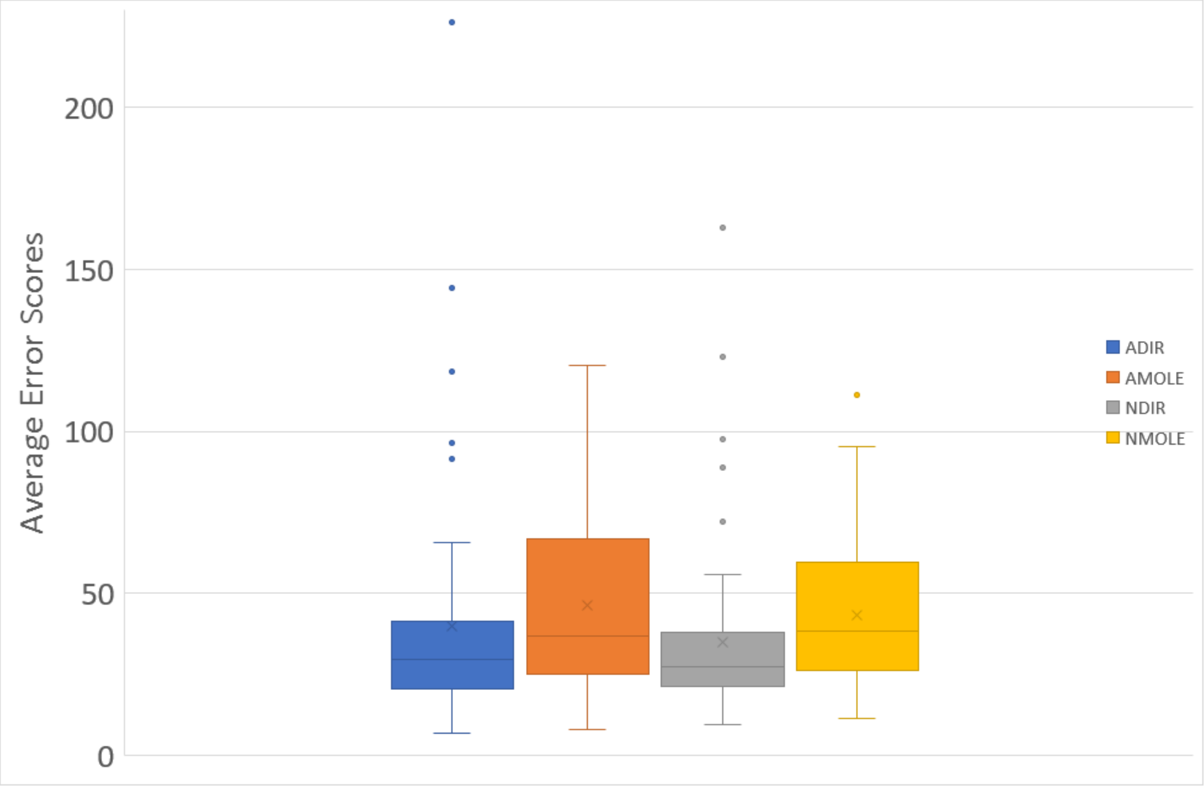


Figure 6 Box and whisker plot of distribution of individual average error scores

3.8.5 Best estimates

The average best estimates for each condition are summarised in a box and whisker plot shown as Figure 8. Although inspection of the chart suggests the presence of outliers, they did not affect results and so were included in analyses. A two-way repeated measures ANOVA was conducted, showing significant effects of both the type of elicitation (F(1,60) = 20.40, $p < .001$, $\eta_p^2 = .25$) and the presence of anchors (F(1,60) = 6.46, $p = .01$, $\eta_p^2 = .10$) on best estimates. A post hoc pairwise t-test with Holm correction found significant differences between the following pairs of conditions: NDIR and NMOLE, $p < .001$; NDIR and ADIR, $p = .03$; NDIR and AMOLE, $p < .001$; ADIR and NMOLE < .001; ADIR and AMOLE, $p < .001$. The only conditions without a significant difference between their best estimates were AMOLE and NMOLE, $p = .61$. These results, together with inspection of the graph demonstrate that AMOLE and NMOLE best estimates were, equally, the highest estimates. Additionally, Figure 8, displays the median best estimates, which for the direct elicitation tasks are close to the actual average of 87.7, with Median(ADIR) = 95 and Median(NDIR) = 88.

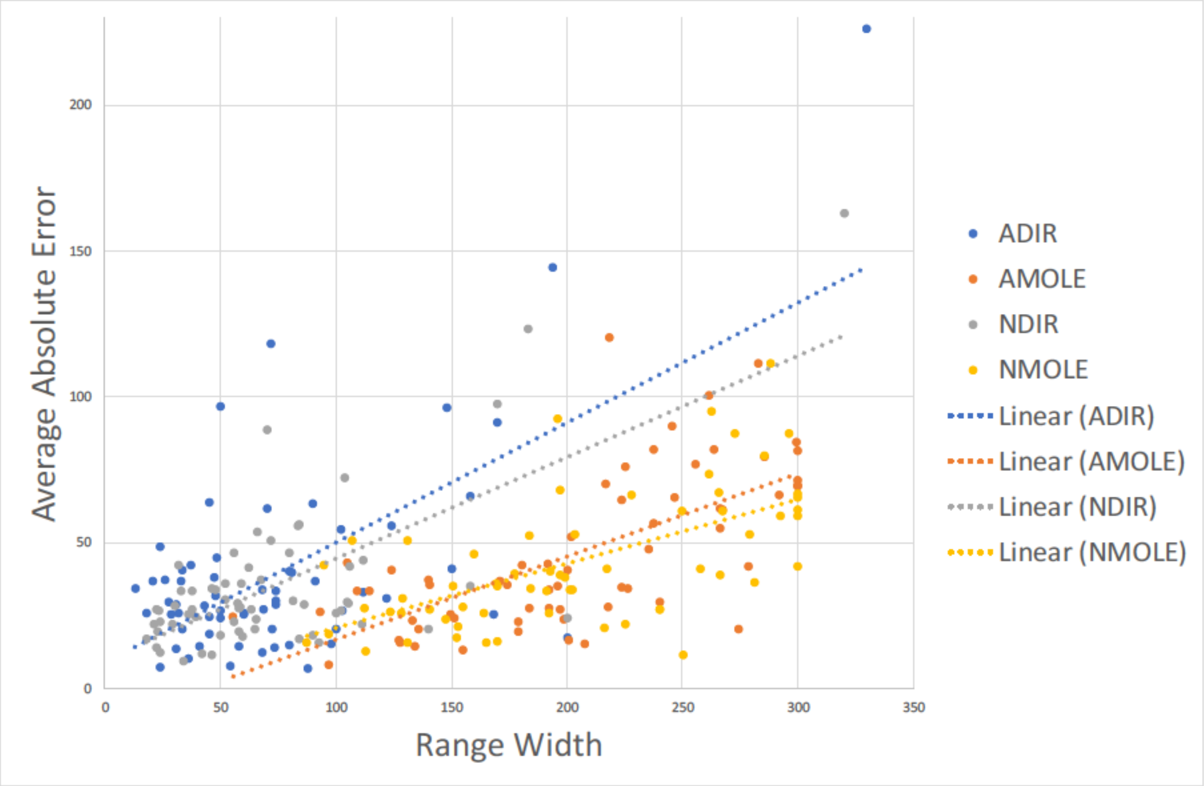


Figure 7 Scatterplot for range width v average error

3.8.6 Order effects

As noted above, ranges were wider in the MOLE conditions and these also contained higher best estimates. From this we can infer that, when participants were presented with high values, they incorporated them into their own range of potential answers. Participants who completed the NDIR condition first did not have any values suggested to them. Their mean range width for the NDIR condition was *M* = 55.64, while the mean NDIR range width for participants who did not compete this condition first was *M* = 79.43. The difference between these values was not quite significant according to a one tailed Welch's independent sample t test (*t*(27) = -1.69, *p* = .051 *d* = 0.49). Performing the same test for the best estimates, a significant difference was found between participants who complete NDIR first (*M* = 88.12) and those who do not (*M* = 103.75) *t*(59) = -2.22, *p* = .015, *d* = 0.50.

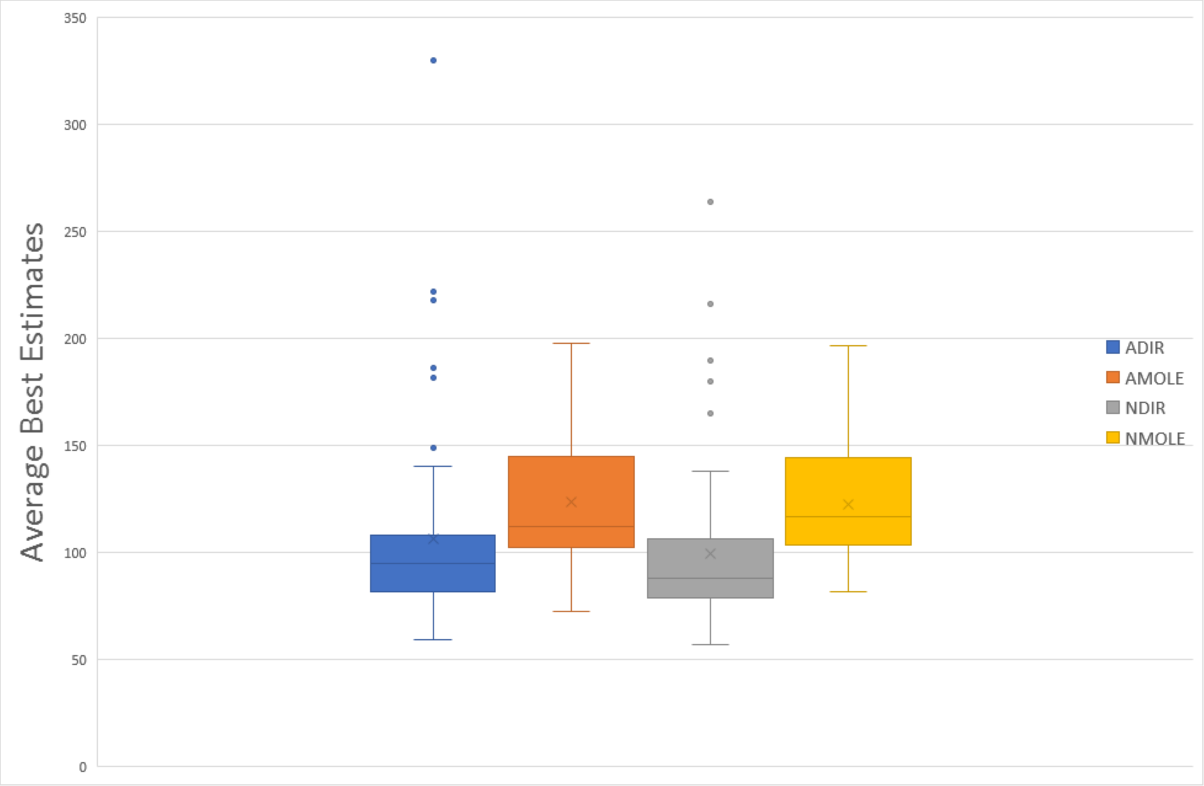


Figure 8 Box and whisker plot for distribution of individual average best estimates

3.8.7 Similarities

The full correlation matrix (Appendix C) indicated that there were several measures that were correlated across conditions. Table 11 shows the correlation of average range widths across conditions, with several significant correlations varying between medium and large effect sizes. In particular, ADIR widths are highly correlated to NDIR widths and AMOLE widths are highly correlated to NMOLE widths. Overall, wider range widths in one condition are likely to be associated with wider range widths in other conditions

Table 12, likewise, displays the correlations for calibration. There is a large positive correlation between the ADIR and NDIR conditions. There are also medium positive correlations between AMOLE and NMOLE, and AMOLE and NDIR conditions. This means that participants who are well calibrated in one condition are also likely to be well calibrated in the other condition within the same elicitation type.

Table 13 shows the correlations for average error. These values are all significantly positively correlated, with medium to large effect size. This indicates that higher errors in any condition are likely to be associated with higher conditions in all the other conditions.

These results all suggest that there are consistencies in the way participants performed across the elicitation tasks – as would be expected given the relationships between ARF engagement and performance noted above.

Table 11 Correlations for average width

Measures	2.	3.	4.
1. ADIR average width	0.20	0.99***	0.36**
2. AMOLE average width	-	0.21	0.80***
3. NDIR average width		-	0.34**
4. NMOLE average width			-

Note: p < .05 = *, *p* <.01 = **, *p* < .001 = ***

Table 12 Correlations for calibration

Measures	**2.**	**3.**	**4.**
1. ADIR calibration	0.17	0.64***	0.22
2. AMOLE calibration	-	0.31*	0.41***
3. NDIR calibration		-	0.13
4. NMOLE calibration			-

Note: p < .05 = *, *p* <.01 = **, *p* < .001 = ***

Table 13 Correlations for average error

Measures	**2.**	**3.**	**4.**
1. ADIR average error	0.46***	0.88***	0.50***
2. AMOLE average error	-	0.54***	0.67***
3. NDIR average error		-	0.54**
4. NMOLE average error			-

Note: p < .05 = *, *p* <.01 = **, *p* < .001 = ***

4 Discussion

This study investigated overconfidence as it relates to More-Or-Less Elicitation (MOLE). The MOLE has been designed to limit the effect of overconfidence and anchoring on elicited values, while also seeking to improve the accuracy of estimates (Welsh & Begg, Under review). The performance of the MOLE was compared with a direct elicitation measures. This information was used to test hypotheses relating to elicitation performance, anchoring, format dependence and individual differences.

4.1 Findings

4.1.1 Elicitation effects

The MOLE has previously demonstrated three main beneficial effects; these effects relate to range width, calibration and accuracy (Welsh & Begg, Under review). The MOLE was compared to direct elicitation in an experimental component of the study; therefore, causal attributions can be made. The direct elicitation comprised of a request for a minimum, a maximum and a best estimate.

The first of the beneficial effects was range widths, with the MOLE producing ranges that were significantly wider than those produced by direct elicitation. This effect can be attributed to the MOLE process, that is, the MOLE causes people to produce wider ranges. The typical assumption is that overconfidence is a result of overly narrow ranges (Mannes & Moore, 2013; Soll & Klayman, 2004; Tversky & Kahneman, 1974). The anticipated benefit of wider ranges is that they will contain the actual value more often. This is exactly what was found in this study and describes the second main effect of interest, calibration. Individual calibration scores indicate how many of the person's ranges contained the actual value. Again, there was significantly better performance for the MOLE than for direct elicitation. In fact, calibrations for the MOLE conditions were 93 and 94%, which amounts to 7 and 6% overconfidence. This is demonstrably better than the 57 and 60% calibrations scores that resulted from direct elicitation, and the 50% calibration (for 80 to 90% confidence intervals) that is commonly reported (Alpert & Raiffa, 1982; Lichtenstein et al., 1982; Soll & Klayman, 2004). Again, we can make a causal attribution and state that the improved calibration was a result of the MOLE.

The third main effect of interest was accuracy. Unlike previous studies, there was no evidence of the MOLE producing more accurate best estimates than direct elicitation.

4.1.2 Anchoring

It has previously been assumed that the MOLE overcomes the effects of anchoring (Welsh & Begg, Under review). To test this assumption, an anchor value was presented prior to direct elicitation tasks and MOLE tasks. A small to medium effect of anchoring was found in the direct elicitation task, however, the MOLE avoided the effect of anchoring under the same conditions. This suggests that the MOLE was able to overcome the effects of anchoring.

4.1.3 Format dependence

This study found evidence of format dependence. Within the direct elicitation tasks, participants produced ranges that they later recognised as being much too narrow. This indicates a format dependence effect, where participants are better calibrated when they evaluate ranges rather than produce ranges. However, in post hoc evaluations, participants were still overconfident; the evaluation errors of section 3.4 show that participants still were 7 and 9% overconfident.

These results were very different for the MOLE conditions. The first notable finding is that the production error was very low (7 and 6%), indicating that the ranges were initially well calibrated. Calibration was not significantly improved by evaluating these ranges. However, participants did seem to indicate that ranges they produced with the MOLE should be wider, based on their underconfident evaluation errors. To summarise, when using the MOLE, participants produced well calibrated ranges, but subsequently assessed them as needing to be wider. This seems unusual, but is consistent with Juslin et al. (1999), who found the same effect of parallel overconfidence and underconfidence when participants completed the same task, but with different elicitation formats.

4.1.4 Individual differences

4.1.4.1 Openness

McElroy and Dowd (2007) found that openness was related to anchoring and this study aimed to replicate these findings. The same openness measure as the original study (TIPI openness subscale) was used, as was a more robust measure of openness (NEO-FFI openness subscale), in an attempt to further validate these findings. However, the current study did not find evidence of correlation between openness and anchoring for either anchoring measure. While anchoring aligns well with the description of openness, the data do not support such a relationship. This is consistent with Furnham et al. (2012), who also failed to replicate the effect.

4.1.4.2 Need for closure

Need for Closure (NFC) was proposed to relate to wider ranges for the MOLE and narrower ranges for direct elicitation. The current study did not replicate the findings of Kaesler et al. (2016) in that range widths were not narrower for higher NFC when direct elicitation was employed. It was also expected that range widths would be wider as NFC increased, for the MOLE conditions. There was some evidence for an effect, albeit small and only significant when considered as a one-tailed test. Overall, these results are not convincing; however, the concept is worthy of further investigation. The NFC measure was included to represent the Informativeness-Accuracy Trade-off (IAT), which aligns with performance observed with the MOLE (see section 4.2.1.3) and, therefore, may be of relevance. The lack of convincing results could be due to three factors. First, the study was underpowered to find an effect of this size. Second, there appears to be range truncation in the NFC measure; while the scale extends from 15 to 90, the scores for participants in this study were grouped between 41 and 80. Third, it also seems that direct elicitation responses were influenced by priming effects (explained in section 4.1.5). All of these factors may have contributed to this study's ability to determine the relationship between NFC and range widths.

4.1.4.3 Conscientiousness

It was anticipated that conscientiousness would have a favourable influence on the care with which people approached participation in this study. However, there was no relationship found between conscientiousness and the measures selected to approximate care taken. The only significant finding of the study that related to conscientiousness was that it was associated with a reduction in anchoring effect in direct elicitation. It is possible that conscientiousness could be associated with identification of unreasonable anchors, although this has not been the case in previous studies with Eroglu and Croxton (2010) finding the opposite effect and McElroy and Dowd (2007) finding no relationship.

4.1.5 Other findings

Engagement is the most pervasive factor in this study. Engagement measured the amount of involvement participants had with engagement with Australian Rules Football and was used as a proxy measure of knowledge. Greater engagement related to improved elicitation performance in four respects. The first was that across all elicitation methods, there was a medium to large negative correlation between knowledge and error; this indicated that greater knowledge was linked to best estimates that were closer to the true values. This is unsurprising and was also found by McKenzie et al. (2008).

The second finding concerning engagement related to range widths. It was expected that participants would adjust their ranges to reflect their knowledge; that is, wider ranges to accommodate less knowledge and more precise ranges afforded by greater knowledge (Block & Harper, 1991; Bruza et al., 2011). This was found to be the case for the MOLE, but not for direct elicitation tasks; the MOLE seems to facilitate the adjustment of range width to accommodate knowledge. The consistently high calibration of the MOLE tasks demonstrates that the range adjustment was, in this case, appropriate in that narrowing of ranges did not tend to result in overconfidence.

The third finding concerning engagement was that the direct elicitation tasks showed moderate correlations between knowledge and calibration, and knowledge and accuracy, but these same results were not found when using the MOLE. The calibration finding is readily explained; in the situation of generally poor calibration found with direct elicitation, knowledge was found to improve responses. Correspondingly, the high level of calibration found with the MOLE did not leave room for improvement based on knowledge. This could also be articulated by saying that the MOLE compensates for low expertise, producing well calibrated ranges, regardless of knowledge. The finding relating to knowledge and accuracy is more troubling; knowledge relates to greater accuracy in direct elicitation, but not the MOLE, which was found to have low to no accuracy. This indicates that not only did the MOLE not produce best estimates that correlated to actual values, but that this did not improve with knowledge.

Finally, the fourth finding was that knowledge was found to relate to reduced anchoring in direct elicitation. This contradicts the findings of Northcraft and Neale (1987) that experts and amateurs were equally affected by anchors. However, this finding makes sense; it would be expected that people who are more knowledgeable would be less affected by extreme and/or unreasonable anchors, as seen in Welsh et al. (2014).

A complicating factor of this study is that the tasks themselves seemed to have a priming effect on participants. The direct elicitation task without anchoring required participants to produce their best estimate, without having any values suggested to them. Participants who completed this task first produced smaller ranges (near significant) and lower best estimates[1] than those participants who completed the same task after being exposed to external numerical cues. The values that are produced by the MOLE are drawn from the initial wide starting range, these values are large due to the range being necessarily skewed by the natural lower bound of zero. This may have had a priming effect on responses. Three possible explanations for a priming effect are, firstly that participants with little knowledge of AFL may have interpreted these values as clues to the correct answers (Furnham & Boo, 2011). Secondly, it is possible that the presented values influenced the answers that participants produced - as regardless of how implausible they are, presented values may increase the accessibility of evidence that supports such values (Mussweiler & Strack, 1999, 2001). Thirdly, magnitude priming could account for a general effect of high values responses occurring when high values are included in the information that is provided to participants (Oppenheimer, LeBoeuf, & Brewer, 2008). Overall, it seems likely that the responses of the direct elicitation tasks were inflated by other aspects of the study.

There were correlations across the elicitation tasks for range widths, calibration and error, indicating that there was consistency of responses. Participants who performed in a particular manner in one condition were likely to perform in the same way in other conditions. Such a pattern of behaviour is indicative of individual differences (McCrae & Costa, 1997). While the individual difference measures utilised in this study did not produce

[1] NDIR first range width M = 55.64, compared to M = 79.43, t(27) = -1.69, p = .051 d = 0.49, one-tailed
NDIR first best estimate (M = 88.12), compared to (M = 103.75), t(59) = -2.22, p = .015, d = 0.50, one-tailed

compelling results, it seems likely that individual differences do play a part in elicitation performance.

4.2 Implications

4.2.1 Cause of overconfidence

The results of the study allow for some insight into the causes of overconfidence.

4.2.1.1 Range widths

This study confirms that overly narrow ranges contribute to overconfidence. Moore et al. (2015) propose that ranges are sufficiently wide, but incorrectly located; however, this was not the case in the current study. The median best estimates produced in the direct elicitation tasks were very close to the mean actual value[2]; therefore, these estimates were appropriately located, however, they were still associated with at least 40% overconfidence, indicating that narrow ranges were, indeed, contributing to overconfidence.

4.2.1.2 Anchoring

While an anchoring effect was identified within this study, it was not found to relate to overconfidence. This is illustrated by comparing the direct elicitation task, both with and without the anchoring task. The presence of an anchor value created no difference in range width or calibration, indicating that anchoring did not cause overconfidence. This is consistent with Block and Harper (1991), who also found anchoring unrelated to overconfidence.

4.2.1.3 Informativeness-Accuracy Trade-Off

The Informativeness-Accuracy Trade-off (IAT; Yaniv & Foster, 1995, 1997) is consistent with the results of this study. When asked to produce ranges through direct elicitation, participants tended to produce narrow ranges, even though they subsequently

[2] ADIR median best estimate = 95, NDIR median best estimate = 88, actual average score = 88

identified that these ranges were inaccurate because they were too narrow. This indicates a preference for being informative, rather than accurate. This manner of selecting ranges is reversed when using the MOLE; rather than selecting values to include in the range, the MOLE asks users to select the values to exclude. Participants were similarly conservative in selecting values to either include or exclude, resulting in narrow ranges in direct elicitation and wide ranges when using the MOLE.

4.2.1.4 Elicitation format

This study supports the premise of format dependence, that overconfident performance can occur as a result of the format used to elicit responses (Winman et al., 2004). Participants were able to assess their own ranges with a higher degree of accuracy, despite being substantially overconfident when they produced them. This indicates that the participants are able to recognise their own overconfidence, and can endorse well calibrated ranges under the right circumstances.

4.2.2 MOLE performance

The findings of this study show that the MOLE is directly responsible for increased calibration, through the primary manipulation of creating wider ranges. The findings do not suggest that the MOLE provided accurate best estimates in this case. While the MOLE appeared to be immune to the effects of anchoring, this offered no benefit in this study, with anchoring not affecting overconfidence. It is possible, however, that under different circumstances the ability to overcome anchoring would be beneficial.

The MOLE is able to address the causes of overconfidence that have been verified in this study. The IAT usually leads to overconfidence, but the MOLE process utilises this tendency to preserve uncertainty to widen ranges. The MOLE takes advantage of format dependence by utilising participants' evaluations to produce well calibrated ranges.

4.3 Caveats

The MOLE is a repetitive task that was performed over an average of 30 minutes. One participant was excluded from the study because they seemed to give up half way through and clicked the buttons as fast as possible to finish. While no other participants displayed such overt impatience with the task, it is anticipated that some participants' results may be suboptimal. While the current study still obtained meaningful results, it is anticipated that the performance of the MOLE can only improve when applied in meaningful settings by enthusiastic or committed users.

There was no requirement for participants to have prior knowledge of AFL. Indeed, 15 of the 62 total participants reported having no or rare involvement with AFL. Interestingly, even with a lack of knowledge, all but one of these participants made decisions about the probability of outcomes; the remaining participant rated all of the random MOLE values as equally likely, which is a genuine response in the lack of any other knowledge. However, the pattern of providing information without appropriate knowledge also strengthens the IAT with participants preferring to provide information, rather than to reply with uncertainty. Restricting the participants to those who had greater knowledge of AFL may have affected results.

Some unusual results were obtained when assessing post hoc confidence in the summary task. It is possible that people failed to read or understand the instructions; the summary task was visually similar to the MOLE task and some participants may have been, incorrectly, trying to indicate the location of their best estimate rather than provide a confidence level. However, some participants also displayed an unusual pattern, they selected lower confidence for wider ranges. It may be that people were reducing their confidence rating to indicate that they believed ranges were too wide. This is consistent with explanations of people not truly understanding probability in mathematical terms, and instead using it to indicate credibility or typicality (Gigerenzer, 2000; Mannes & Moore, 2013).

4.4 Future Research

Despite a lack of compelling results for the individual difference measures, it is recommended that further research be conducted in this area. The patterns of response that were seen in this study suggest involvement of personality traits. NFC is worth further investigation as the production of significant results seemed to have been hampered by range truncation, a lack of power and possible priming effects. In general, the study was underpowered to find small effects, which are typically found in decision making (Welsh et al., 2014); therefore, future research should include a priori power analysis to determine appropriate sample sizes.

The priming effect may have increased elicited values throughout the study, this could be limited by a between participants design, with separate pools of participants completing direct elicitation and the MOLE. This would ensure that MOLE values did not have a magnitude priming effect on direct elicitation measures.

Future studies should investigate the manner in which the MOLE calculates best estimates. The poor performance of the MOLE in producing best estimates was unexpected and investigating the particular conditions that led to this performance provides the opportunity to optimise the MOLE algorithms.

4.5 Conclusion

The results of this study indicate that the MOLE is an effective tool for reducing overconfidence and avoiding the effect of anchoring values. However, in this case, there was poor performance when producing a best estimate. Individual difference measures were substantially unrelated to either overconfidence and anchoring. The study was, however, able to provide some insight as to the cause of overconfidence: narrow ranges widths, IAT and format dependence were all found to contribute to overconfidence. Anchoring was not found to cause overconfidence, contradicting the original explanation of overconfidence arising from the anchoring and adjustment heuristic (Tversky & Kahneman, 1974).

5 References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probablility assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294-305). Cambridge: Cambridge University Press.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*(456). doi:10.3389/fpsyg.2017.00456

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49*(2), 188-207. doi:http://dx.doi.org/10.1016/0749-5978(91)90048-X

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, Boston, USA.

Costa, P. T., Jr., & McCrae, R. R. (2003). NEO Five-Factor Inventory test booklet form S (Adult). Lutz, Florica, USA: Psychological Assessment Resources, Inc.

Costa, P. T., Jr., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The Sage Handbook of Personality Theory and Assessment* (Vol. 2 Personality Measurement and Testing, pp. 179-198). London: Sage.

Eroglu, C., & Croxton, K. L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting, 26*(1), 116-133. doi:10.1016/j.ijforecast.2009.02.005

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

Fischhoff, B. (1988). Judgment and decision making. In R. J. Sternberg & E. E. Smith (Eds.), *The Psychology of Human Thought* (pp. 153-187). Cambridge: Cambridge University Press.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*(1), 35-42. doi:10.1016/j.socec.2010.10.008

Furnham, A., Boo, H. C., & McClelland, A. (2012). Individual differences and the susceptibility to the influence of anchoring cues. *Journal of Individual Differences, 33*(2), 89-93. doi:10.1027/1614-0001/a000076

Gigerenzer, G. (2000). Adaptive thinking : Rationality in the real world (pp. 241-266). New York, UNITED STATES: Oxford University Press.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Goldberg, L. R. (1992). The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment, 4*(1), 26-42.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528. doi:10.1016/s0092-6566(03)00046-1

Grice, H. P. (1975). Logic and conversation. In J. P. Kimball (Ed.), *Syntax and Semantics* (pp. 41-58): Academic Press.

Hawkins, J. T., Coopersmith, E. M., & Cunningham, P. C. (2002). *Improving stochastic evaluations using objective data analysis and expert interviewing techniques*. Paper presented at the Society of Petroleum Engineers 78th Annual Technical Conference and Exhibition, San Antonio, Texas.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences, 18*(10), 504-506. doi:https://doi.org/10.1016/j.tics.2014.06.009

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format Dependence in Subjective Probability Calibration. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*(4), 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), 678-703. doi:10.1037/0033-295X.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* (1st ed.). New York: Farrar, Straus and Giroux.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge ; New York: Cambridge University Press.

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 Universtiy Press.

Mannes, A. E., & Moore, D. A. (2013). A Behavioral Demonstration of Overconfidence in Judgment. *Psychological Science, 24*(7), 1190-1197. doi:10.1177/0956797612470700

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality Trait Structure as a Human Universal. *American Psychologist, 52*(5), 509-516.

McElroy, T., & Dowd, K. (2007). Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and Decision Making Journal, 2*(1), 48-53.

McKenzie, C. R. M., 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. doi:https://doi.org/10.1016/j.obhdp.2008.02.007

Moore, D. A., Carter, A. B., & Yang, H. H. J. (2015). Wide of the mark: Evidence on the underlying causes of overprecision in judgment. *Organizational Behavior and Human Decision Processes, 131*, 110-120. doi:http://dx.doi.org/10.1016/j.obhdp.2015.09.003

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 15*(2), 502-517.

Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence? *Personality and Individual Differences, 37*(5), 1013-1022. doi:10.1016/j.paid.2003.11.010

Mussweiler, T., & Strack, F. (1999). Hypothesis-Consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*(2), 136-164. doi:https://doi.org/10.1006/jesp.1998.1364

Mussweiler, T., & Strack, F. (2001). Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition, 19*(2), 145-160. doi:10.1521/soco.19.2.145.20705

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. doi:Doi 10.1177/01461672002611010

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*(1), 84-97. doi:https://doi.org/10.1016/0749-5978(87)90046-X

Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition, 106*(1), 13-26. doi:https://doi.org/10.1016/j.cognition.2006.12.008

Rammstedt, B., Danner, D., & Martin, S. (2016). The association between personality and cognitive ability: Going beyond simple effects. *Journal of Research in Personality, 62*(Supplement C), 39-44. doi:https://doi.org/10.1016/j.jrp.2016.03.005

Roets, A., & Van Hiel, A. (2007). Separating ability from need: Clarifying the dimensional structure of the Need for Closure Scale. *Personality and Social Psychology Bulletin, 33*(2), 266-280. doi:doi:10.1177/0146167206294744

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences, 50*(1), 90-94. doi:http://dx.doi.org/10.1016/j.paid.2010.09.004

Russo, J. E., & Schoemaker, P. J., H,. (1992). Managing overconfidence. *Sloan Management Review, 33*(2), 7-17.

Schaefer, P. S., Williams, C. C., Goodie, A. S., & Campbell, W. K. (2004). Overconfidence and the Big Five. *Journal of Research in Personality, 38*(5), 473-480. doi:https://doi.org/10.1016/j.jrp.2003.09.010

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

Tukey, J. W. (1977). *Exploratory data analysis*: Addison-Wesley Publishing Company.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Webster, D. M., & Kruglanski, A. W. (1994). Individual-Differences in Need for Cognitive Closure. *Journal of Personality and Social Psychology, 67*(6), 1049-1062. doi:Doi 10.1037/0022-3514.67.6.1049

Welsh, M. B., & Begg, S. H. (2015). *Reducing overconfidence in forecasting with repeated judgement elicitation*. Paper presented at the 37th Annual Meeting of the Cognitive Science Society, Austin, Texas.

Welsh, M. B., & Begg, S. H. (2016). What have we learnt? Insights fromm a decade of bias research. *Australian Petroleum Production and Exploration Association Journal, 56*, 435-450.

Welsh, M. B., & Begg, S. H. (Under review). More-Or-Less Elicitation (MOLE): Reducing bias in range estimation and forecasting. *EURO Journal of Decision Processes*.

Welsh, M. B., Delfabbro, P. H., Burns, N. R., & Begg, S. H. (2014). Individual differences in anchoring: Traits and experience. *Learning and Individual Differences, 29*, 131-140. doi:10.1016/j.lindif.2013.01.002

Welsh, M. B., Navarro, D. J., & Begg, S. H. (2011). *Number preference, precision and implicity confidence*. Paper presented at the Annual Conference of the Cognitive Science Society (CogSci), Boston.

Welsh, M. B., Lee, M. D., & Begg, S. H. (2008). *More-or-Less Elicitation (MOLE): Testing a heuristic elicitation method*. Paper presented at the Annual Conference of the Congitive Science Society (CogSci), Washington DC.

Welsh, M. B., Lee, M. D., & Begg, S. H. (2009). *Repeated judgements in elicitation tasks: Efficacy of the MOLE method*. Paper presented at the Annual Conference of the Cognitive Science Society (CogSci), Amsterdam.

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology. Learning, Memory & Cognition, 30*(6), 1167-1175. doi:10.1037/0278-7393.30.6.1167

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, .124*(4), pp. doi:10.1037/0096-3445.124.4.424

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making, .10*(1), pp. doi:10.1002/%28SICI%291099-0771%28199703%2910:1%3C21::AID-B DM243%3E3.0.CO;2-G

Appendix A

RESEARCH PARTICIPANTS NEEDED

You are Invited…

to participate in a study that investigates how accurately people can forecast common events. You will be asked to predict the results for future AFL matches.

The study involves a computer based survey that will take less than 45 minutes to complete.

Participants will receive a $20 Coles Myer gift card.

If you are interested, please contact:

✉

[redacted]

☎

[redacted]

To be eligible:

You must be 18 years of age or older

You must be able to read and write in English

You must be currently living in Australia

You must be able to collect your gift card from The University of Adelaide

This project is being conducted by Marianne Clausen. This research will form the basis for the degree of Bachelor of Psychological Science Honours at the University of Adelaide under the supervision of Dr Matthew Welsh. The study has been approved by the Psychology Sub-Committee of the Human Research Ethics Committee at the University of Adelaide (approval number H-2017-17/46).

Appendix B

MOLE AFL Study

Only available in MS word version. Double click to open the program. Enable macros.

Email [redacted] for a copy if embedded document unavailable.

Appendix C

Table C 1 Correlation matrix of continuous variables

	Engage AFL	NFC	TIPI_O	NEOFFI_O	NEOFFI_C	ADIR ACCURACY	ADIR ANCHORING	AMOLE ACCURACY	AMOLE ANCHORING	NDIR ACCURACY	NMOLE ACCURACY	ADIR_AVWIDTH	AMOLE_AVWIDTH	NDIR_AVWIDTH	NMOLE_AVWIDTH	ADIR_CAL	AMOLE_CAL	NDIR_CAL	NMOLE_CAL	ADIR_AVERR	AMOLE_AVERR	NDIR_AVERR	NMOLE_AVERR	
Age	0.15	-0.057	0.04	0.05	0.10	0.42***	-0.20	-0.07	-0.05	-0.02	0.04	-0.16	-0.17	-0.10	-0 21	0.11	0.08	0.20	-0.07	-0.30*	-0 22	-0.24	-0 31*	Age
Engage AFL	1	-0.270	0.17	-0.04	0.14	0.44***	-0.46***	0.14	-0.15	0.40**	-0.20	-0.16	-0 53***	-0.13	-0.58***	0 35**	-0.04	0.40**	-0.03	-0.41***	-0.60***	-0.43***	-0.47***	Engage AFL
NFC		1	-0.45***	-0.14	0.16	-0.14	-0.12	0.00	0.05	0.06	0.03	0.01	0.21	-0.03	0 24	0.10	0.16	0.01	0.12	0.05	0.13	0.01	0.19	NFCC1
TIPI O			1	0.49***	-0.13	0 29*	-0.03	0.20	-0.16	0.11	-0.08	-0.04	-0.05	-0.05	-0.09	-0.07	-0.14	-0.07	-0.15	-0.13	0.04	-0.07	0.01	TIPI1
NEOFFI O				1	-0.11	0.11	-0.07	0.10	-0.12	-0.03	-0.13	-0.04	0.02	-0.04	-0.04	0.08	-0.17	-0.01	0.03	-0.16	0.06	-0.05	0.07	NEOFFI O
NEOFFI_C					1	0.07	-0.30*	0.10	-0.01	0.01	0.14	-0.15	-0.03	-0.20	-0.09	0.10	-0.24	0.01	-0 24	-0.20	-0.03	-0.12	0.09	NEOFFI_C
ADIR ACCURACY						1	-0.33**	0.25*	-0 32*	0.20	-0.06	-0.12	-0.13	-0.08	-0 21	0 25*	-0.02	0.24	-0.04	-0.34**	-0.15	-0.19	-0.06	ADIR ACCURACY
ADIR ANCHORING							1	-0.17	0.13	-0.14	0.21	0.33**	0.10	0.27*	0 28*	-0.24	-0.04	-0.18	0.03	0.44***	0 30*	0.31	0 21	ADIR ANCHORING
AMOLE ACCURACY								1	-0.13	-0.15	0.02	-0.11	-0.07	-0.17	-0.02	0.11	0.08	0.03	-0.07	-0.19	-0 24	-0.14	0.16	AMOLE ACCURACY
AMOLE ANCHORING									1	-0.15	0.02	0.03	0.06	-0.04	0.11	-0.12	0.00	-0.10	0 22	0.12	0.00	0.04	0.01	AMOLE ANCHORING
NDIR ACCURACY										1	-0.18	-0.14	-0.25	-0.12	- 30*	0.19	-0.16	0.21	-0.10	-0.21	-0.08	-0.36**	-0.09	NDIR ACCURACY
NMOLE ACCURACY											1	0.15	0.18	0.15	0.17	-0.09	0.08	-0.17	-0.05	0.15	0.16	0.23	-0.04	NMOLE ACCURACY
ADIR AVWIDTH												1	0.20	0.99***	0.36**	0 29*	0.05	0.26*	0 21	0.64***	0 30*	0.64	0 35	ADIR AVWIDTH
AMOLE_AVWIDTH													1	0.21	0.80***	-0.06	0.44***	-0.21	0 27	0.19	0.67***	0.33	0.45	AMOLE_AVWIDTH
NDIR AVWIDTH														1	0.34**	0.24	0.09	0.33**	0.16	0.64***	0 26	0.68***	0 29	NDIR AVWIDTH
NMOLE AVWIDTH															1	-0.09	0.24	-0.26*	0 37**	0.41***	0.60	0.48	0.61***	NMOLE AVWIDTH
ADIR CAL																1	0.17	0.64***	0 22	-0.42***	-0 28	-0.27	-0.17	ADIR CAL
AMOLE CAL																	1	0.31*	0.41***	-0.11	-0.17	-0.12	-0 21	AMOLE CAL
NDIR_CAL																		1	0.13	-0.26*	-0.45	-0.30*	-0 34	NDIR_CAL
NMOLE CAL																			1	0.06	-0.02	0.05	-0.10	NMOLE CAL
ADIR AVERR																				1	0.46***	0.88***	0.50***	ADIR AVERR
AMOLE AVERR																					1	0.54***	0.67***	AMOLE AVERR
NDIR AVERR																						1	0 54**	NDIR AVERR
NMOLE_AVERR																							1	NMOLE_AVERR