Does anchoring cause overconfidence only in experts?

Belinda Bruza, Matthew B. Welsh, Daniel J. Navarro & Stephen H. Begg
({belinda.bruza, matthew.welsh, daniel.navarro, steve.begg}@adelaide.edu.au)
The University of Adelaide, SA 5005, Australia

Abstract

The anchoring-and-adjustment heuristic (Tversky & Kahneman, 1974) predicts elicitation of an initial estimate will prompt subsequent minimum and maximum estimates to lie close to the initial estimate, resulting in narrow ranges and overconfidence. Evidence for this, however, is mixed; while Heywood-Smith, Welsh & Begg (2008) observed narrower subsequent ranges, Block and Harper (1991) report ranges became wider. One suggestion has been that this reflects a difference between expert and novice reactions to elicitation tasks. The present study investigated whether the interplay between expertise and number preferences leads to the paradoxical effects of an initial estimate. Participants with high expertise make precise estimates whereas participants with less expertise prefer rounded numbers, which could, potentially, reduce the impact of anchors. We confirm that expertise affects the precision of estimates and observe results indicative of the theorized effect – an interaction between expertise and elicitation method on range widths.

Keywords: anchoring; overconfidence; number preference; precision

In fields where empirical data is limited or unavailable, decisions are often based on expert judgment. For example, current industry practice in petroleum exploration requires exploration geologists to provide 80% confidence ranges on relevant factors (e.g., rock porosity, reservoir thickness) prior to drilling (Hawkins, Coopersmith, & Cunningham, 2002). A typical result, however, is overconfidence (Lichtenstein, Fischhoff, & Phillips, 1982), where the level of confidence reported is much higher than the proportion of ranges containing the true value. This bias has been observed not only in oil and gas industry personnel (Welsh, Bratvold, & Begg, 2005), but in a multiplicity of experts including clinicians (Christensen-Szalanski & Bushyhead, 1981), business managers (Russo & Schoemaker, 1992) and social scientists (Tetlock, 1999). Theoretical interest in factors affecting overconfidence is therefore shared by technical and psychological disciplines alike.

A popular explanation for overconfidence stems from the anchoring-and-adjustment heuristic, first suggested by Tversky and Kahneman (1974): people start from an initial value, an anchor, which they insufficiently adjust from to provide a range. While this anchoring-and-adjustment explanation has received support (Russo & Schoemaker, 1992; Heywood-Smith, Welsh & Begg 2008), several studies found that requesting a best initial estimate resulted in wider ranges, that is, reduced overconfidence (see, e.g., Block & Harper, 1991; Clemen 2001; Juslin, Wennerholm and Olsson, 1999; Soll & Klayman, 2004; Winman, Hansson, & Juslin, 2004).

Yaniv and Foster (1995) theorized there is a trade-off between accuracy and informativeness in uncertain judgment tasks. The precision or "graininess" in estimates is used to convey confidence. On the aforementioned calibration task, for example, an individual uncertain of their knowledge should produce a wide, less precise range to represent uncertainty. However, although wider ranges are more likely to encompass the true value, as estimates become less precise (i.e., "grainier"), they also become less informative of the true value.

There is a possibility that, in order to boost informativeness, experts in a topic are more inclined to generate precise estimates than laypeople. Should this indeed be the case, such a difference in number preference may help clarify the relationship between anchoring and overconfidence.

Such number preferences could place limits on the minimum width of a range that vary by elicitation method. For example, an individual who prefers to give estimates in multiples of 100 (to characterize their uncertainty about the true values) may generate a range of 100-200. If requested to provide an initial best guess, using the same scale this person would estimate either 100 (prompting a wider range of 0-200) or 200 (range: 100-300). The wider range resulting from this preference for round numbers would therefore remove any anchoring effect the initial best guess had on the end-points (and, thereby, reduce overconfidence). Where uncertainty is high and precision low, this effect may be sufficient to overwhelm any anchoring effect resulting from the best guess. In contrast, an expert's tendency to produce precise estimates (i.e., fewer trailing zeros) will reduce or avoid this effect and thus any effect of anchoring resulting from the best guess will be observable.

Research Aims

The aim of this study is to investigate the effect an initial best guess of a true value has on the width of elicited ranges at different gradations of expertise. It was hypothesized that individuals with less expertise would prefer to report estimates in rounded numbers. A best guess would be made as, for example, a multiple of 10. Subsequent adjustment from this anchor would be made on the same scale to obtain minimum and maximum estimates, thereby reducing the impact of anchoring. Conversely, highly expert individuals would report precise estimates. Anchoring on the best guess would therefore be more apparent as adjustments for ranges are made on a smaller scale.

Method

Participants

Participants were 307 undergraduate psychology students studying at the University of Adelaide (83 males and 224 females), aged 16 to 53 years (*M* = 20.07, *SD* = 4.68) who participated for course credit.

Materials

Two purpose-designed 20-item questionnaires were used to assess number preference and the effect of an initial best guess at different gradations of self-rated expertise. The questionnaires comprised Australian Football League (AFL) and general knowledge trivia. There were two experimental conditions – best guess first and range only. For example, on the AFL trivia item *"In what year did the Adelaide Crows join the AFL?"*; participants in the best guess first condition would provide their best estimate of the actual answer before a range (i.e., a low and a high guess) which they were 80% confident contained the actual answer. Participants in the range only condition did not provide an initial best guess. In addition to these confidence intervals, participants rated their confidence that their answer contained the true value, on a 3-point scale: 1 (Absolutely no idea), 2 (I had a vague idea) and 3 (I felt that I knew). Confidence was assessed as the average of all confidence ratings across questions.

Procedure

Data was collected online using SurveyMonkey. In addition to demographics (age and gender), participants were asked to self-rate their expertise: *"What percentage of the Australian population do you have more knowledge of AFL than?"*

Participants were also asked about their engagement in football-related activities, i.e., *"How many AFL games do you watch per week?"*; and *"How many years have you been following AFL?"* Other questions were scored on a Likert scale: *"Do you play football?"* (0 = No; 1 = Yes); *"How often do you attend AFL games?"* (0 = *Never* to 5 = *Weekly*); *"How often do you read or watch news reports about football?"* (0 = *Never* to 4 = *Daily*).

Allocation to one of the two conditions (best guess first or range only) was randomized, but all participants completed the AFL questionnaire before the general knowledge questionnaire.

Results

Scoring

Range To enable comparisons across questions with answers of varying magnitudes, the distance between minimum and maximum estimates on each question was recorded as the *relative range* –the maximum minus the minimum estimate, divided by the true value. Higher scores indicated wider ranges.

Precision Number preference was assessed in terms of *precision* – the number of final zeros in an estimate. For example, an estimate of 100 (2 final zeros), would be scored at precision 2. Lower scores therefore indicated greater precision.

Error As our error measure we used *proportional error.* This was calculated as the average of all error scores proportional to the true value. For the range only condition, error was assessed as the absolute difference between the midpoint of the participant's provided range and the true answer. For the best guess first condition, error was measured as the absolute difference between their best guess and the true answer for each question. Thus, higher scores denoted greater error.

Preliminary Analyses

Preliminary analyses were conducted to ensure expertise on the AFL questionnaire was appropriately measured by self-ratings.

Spearman rank order correlations confirmed self-rated AFL expertise correlated positively with football-related activities. The number of games watched weekly (ρ = .54), years individuals followed AFL (ρ = .54), reading or watching AFL news (ρ = .46) and AFL game attendance (ρ = .42) all had moderate correlations with self-rated expertise (all p < .001). The correlation between actually playing football and self-rated expertise was weak (ρ = .19, p = .001).

Looking at correlations between self-rated AFL expertise and error on each of the AFL trivia questions in the range only condition, 18 of 20 reached significance in the predicted negative direction, ranging from ρ = -.15, p = .03 to ρ = -.46, p < .001. Only one correlation between self-rated AFL expertise and error was positive, ρ = .23, p = <.01.

Similarly, in the best guess first condition, 18 of the 20 correlations between self-rated AFL expertise and error reached significance in the predicted negative direction, ranging from ρ = -.18, p = .03 to ρ = -.47, p < .001. The same item produced a positive correlation between self-rated AFL expertise and error, ρ = .27, p = <.01.

A non-parametric one-tailed sign test indicates the overall negative trend (i.e., 18 out of 20 correlations in the negative direction) is, itself, significant, $p = 2.0\text{x}10^{-4}$.

Mean correlations between AFL expertise and error in the range only and best guess conditions were ρ = -.15, p < .001 and ρ = -.19, p < .001, respectively.

Table 1 shows that confidence (i.e., the average of all confidence ratings reported in the questionnaire) had a moderate, positive correlation with self-rated expertise (ρ = .56, p <.001). The correlation between confidence and error (ρ = -.72, p <.001) was higher than the correlation between self-rated expertise and error (ρ = -.47, p <.001). Confidence was therefore used to indicate expertise on the general knowledge questionnaire. The correlation between confidence and error on the general knowledge task was weaker but in the predicted direction (ρ = -.31, p <.001; see Table 2).

Table 1: Spearman correlation matrix for AFL questionnaire variables

	1	2	3	4	5
1 Expertise	-	<.001	<.01	<.001	<.001
2 Conf.	.56	-	<.01	<.001	<.001
3 Precision	-.16	-.16	-	<.001	.14
4 Range	-.42	-.58	.46	-	<.001
5 Error	-.47	-.72	.07	.56	-

Note: Lower triangle cells show the correlation ρ. Upper triangle cells show the p-value. N = 263. Precision, range and error are averages across questions.

Table 2: Spearman correlation matrix for general knowledge questionnaire variables

	1	2	3	4
1 Conf.	-	<.001	<.001	<.001
2 Precision	-.31	-	<.001	<.001
3 Range	-.21	.58	-	<.001
4 Error	-.31	.24	.47	-

Note: Lower triangle cells show the correlation ρ. Upper triangle cells show the p-value. N = 280. Precision, range and error are averages across questions.

Defining Expertise

Expertise in AFL Self-ratings of AFL expertise were split such that participants rating their knowledge as less than that of 50% of the Australian population were grouped 'low expertise'. Remaining participants who rated their knowledge as greater than or equal to 50% were grouped 'high expertise'.

Expertise in General Knowledge Participants who reported an average confidence rating of less than 2 were grouped 'low expertise'. Remaining participants with an average confidence rating greater than or equal to 2 were 'high expertise'.

Interactions between Expertise and Elicitation Method

It was hypothesized that eliciting a best guess first would cause observable anchoring in high expertise participants; while a anchoring effect in low expertise participants could prompt a greater widening of range end-points.

Figure 1 shows that, on the AFL questionnaire, best guesses led to wider ranges in both expertise groups and high expertise participants gave narrower ranges[1] (range only tM_{20} = .041, CI_{95} = .030, .053; best guess first tM_{20} = .074, CI_{95} = .056, .096) than low expertise participants (range only tM_{20} = .110, CI_{95} = .097, .124; best guess first tM_{20} = .176, CI_{95} = .155, .198).

The same pattern was found for expertise and condition on the general knowledge questionnaire: the best guess first condition produced wider ranges and participants with high expertise had narrower mean ranges (range only tM_{20} = .109, CI_{95} = .092, .128; best guess first tM_{20} = .150, CI_{95} = .125, .178) than participants with low expertise (range only tM_{20} = .157, CI_{95} = .145, .170; best guess first tM_{20} = .265, CI_{95} = .241, .290).

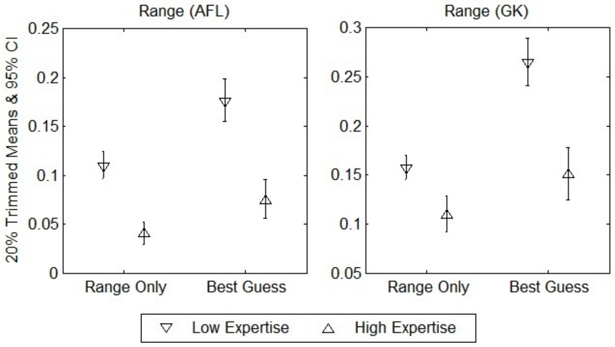

Figure 1: 20% trimmed mean range and 95% confidence intervals for low and high expertise participants in range only (RO) and Best guess first (BG) conditions of the AFL questionnaire (left) and general knowledge questionnaire (right). AFL low expertise RO N = 105; BG N = 82. High expertise RO N = 45; BG N = 31. General knowledge low expertise RO N = 116; BG N = 84. High expertise RO N = 47; BG N = 33.

Visual inspection of the pattern of results is suggestive of an interaction effect of expertise on condition on both AFL and general knowledge questionnaires: that is, the results suggest that the ranges given by low expertise people are being more strongly affected by the inclusion of a best guess than those of experts. Standard two-way analyses of variance[2], however, indicated these interactions were not significant on neither the AFL (F(1, 259) = .29, p = .59,

[1] Variables violated the assumptions of standard parametric procedures; therefore 20% trimmed means are reported to improve robustness against outliers and skewness (Keselman, Algina, Lix, Wilcox, & Deering, 2008). Confidence intervals around these means were calculated using a percentile bootstrap method with 10,000 bootstrap samples (see Erceg-Hurn & Mirosevich, 2008).

[2] Because data was skewed, a rank transformation was performed on all observations for the range of estimates, with the lowest rank of "1" assigned to the smallest observation (see Conover & Iman, 1981).

partial η^2= .001), nor the general knowledge questionnaire (F(1, 276) = .04, p = .84, partial η^2 = .00). It is worth noting that Levene's test on range in the general knowledge questionnaire indicated the assumption of homogeneity of variance was not met (F(3, 276) = 7.57, p < .001).

Given that the ANOVAs checking for these interactions were conducted on the ranks for range, there are also concerns regarding the statistical power of the test, particularly as there is a further loss of power in the ANOVA result for the general knowledge task resulting from the combination of unequal variances with uneven sample sizes. As a result, the reliability of the ANOVA results can be questioned.

As a result of this and the direct observations of Figure 1, which seem to imply an interaction effect of noticeable strength, we conducted an additional analysis.

Testing for an interaction effect between expertise and elicitation method is non-trivial in this case. This is because we wish to test the interaction on the 20% trimmed means (not the mean or median), controlling for possible main effects, without assuming normality. To do so, we constructed a nonparametric permutation-based test. Our test statistic was the extent to which the cell-20% trimmed means deviated from the values predicted by a model consisting solely of main effects (the extent of this variation is formalized via the standard deviation). The distribution of this statistic under the null hypothesis is estimated by constructing 100,000 random permutations of the grouping variables (i.e., elicitation method and expertise status). The p-value is estimated as the probability of observing a deviation from the main effect model predictions as large as or larger than the observed value. For the AFL data, the observed value of .034 is highly significant relative to the null distribution that has mean .01 and std. dev .005 (p < .001). For the general knowledge data, we obtained a test statistic of .04, evaluated against a null distribution with mean .01 and std. dev .006 (p < .001).

Main Effect of Precision

Figure 2 confirms the prediction that high expertise participants would produce more precise estimates (range only tM_{20} = .168, CI_{95} = .133, .204; best guess first tM_{20} = .165, CI_{95} = .124, .208) than low expertise participants (range only tM_{20} = .282, CI_{95} = .258, .306; best guess first tM_{20} = .301, CI_{95} = .274, .329) on the AFL questionnaire.

High expertise participants also provided more precise estimates (range only tM_{20} = .433, CI_{95} = .395, .471; best guess first tM_{20} = .492, CI_{95} = .448, .537) than less expert participants (range only tM_{20} = .560, CI_{95} = .536, .584; best guess first tM_{20} = .586, CI_{95} = .558, .613) on the general knowledge items.

Additional Findings

A main effect of precision on condition was found for high expertise on the general knowledge questionnaire: estimates were more precise in the range only condition (tM_{20}= .433,CI_{95} = .395, .471; best guess first tM_{20}= .492, CI_{95} = .448, .537; see Figure 2).

As depicted in Figure 3, on general knowledge items, high expertise participants produced less error (range only tM_{20} = 8.830, CI_{95} = 6.979, 11.036; best guess first tM_{20} = 10.647, CI_{95} = 8.280, 13.379) than participants with low expertise (range only tM_{20} = 17.448, CI_{95} = 15.602, 19.521; best guess first tM_{20} = 15.578, CI_{95} = 13.879, 17.648).

On the AFL questionnaire, participants with high expertise (range only tM_{20} = 5.935, CI_{95} = 4.041, 9.362; best guess first tM_{20} = 2.711, CI_{95} = 1.685, 4.712) had less error than low expertise participants in the best guess first condition only (range only tM_{20} = 9.039, CI_{95} = 7.558, 11.903; best guess first tM_{20} = 6.647, CI_{95} = 5.298, 8.558).

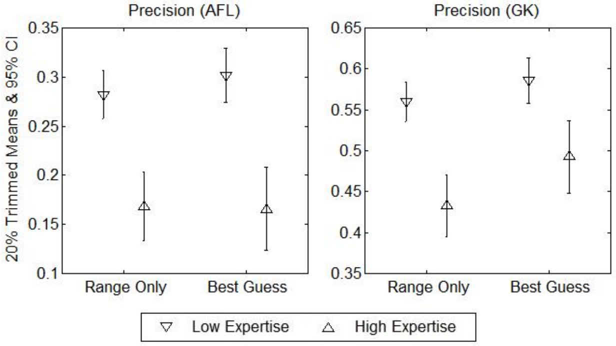

Figure 2: 20% trimmed mean precision and 95% confidence intervals for low and high expertise participants in range only and best guess first conditions of the AFL questionnaire (left) and general knowledge questionnaire (right). Sample sizes as in Figure 1.

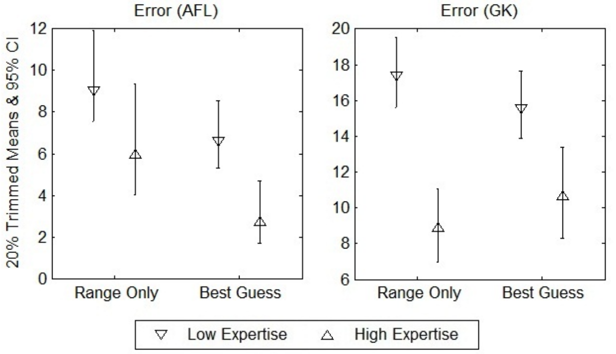

Figure 3: 20% trimmed mean error and 95% confidence intervals for low and high expertise participants in range only and best guess first conditions of the AFL questionnaire (left) and general knowledge questionnaire (right). Sample sizes as in Figure 1.

Note on Analyses

It is important to note that although expertise was discretized in the above analyses, preliminary linear regression analyses[3] with full continuous variables showed the same pattern of results.

Discussion

The results of this study showed main effects of both expertise and elicitation method (group). Participants with a high level of expertise reported estimates with greater precision than participants with less expertise and, in both cases, people who were asked for a best estimate first tended to give wider ranges.

The most interesting result, however, is the interaction between these two. While difficult to analyze, due to violations of the assumptions of parametric tests and the accompanying loss of power in alternative tests, our interpretation of the data, both visually, from Figure 1 and statistically, using a specifically designed permutation test, lead us to conclude that people with low expertise were disproportionately affected by the inclusion of a best guess in both the AFL and general knowledge questions.

That is, less expert people, when asked to estimate a range after having their best guess elicited, increase the width of those ranges more than do more expert people.

This, we argue, may result from their greater preference for rounded numbers, which causes a sort of 'buffering' effect, whereby people's estimates are forced wider because their best guess is already occupying one of the numbers that they would otherwise have used as the end-point of their range.

Caveats

However, a number of caveats should be taken into account when considering our results, including the difficulties we have encountered in analyzing the data. Traditional, parametric tests fail to yield reliable results when their assumptions are violated, yet their non-parametric equivalents often result in a loss of power – which makes the observation of interaction effects particularly difficult. This has necessitated our creation of a specific test for the interaction that we could see in Figure 1.

Other concerns relate to the degree of expertise and number preference observed in our data. Less than a third of our sample rated themselves as better than 50% of the population in the AFL questions and confidence was lower on the general knowledge questions. With a mean self-rated expertise of less than 30%, our sample may, as a result suffer from restricted range, which would undermine the strength of any observed effects. The fact that expertise was self-rated and correlated with the other measures less well than a 3-point confidence rating also suggests that our division between high and low expertise may be more arbitrary than we would hope.

Similarly, the degree of number preference shown on the AFL task, in particular, is extremely low, with the group averages ranging from .075 to .2 – indicating that, at most, people used an extra zero on every fifth estimate. This is much lower than rates observed in other experiments (see, e.g., Welsh, Navarro & Begg, submitted, where an equivalent value above .9 was observed).

Given this it could, reasonably, be argued that our experiment underestimates the magnitude of differences between experts and non-experts – particularly on tasks where uncertainty is higher.

This may also explain the observation that both our 'expert' and 'non-expert' groups widened their ranges as a result of the inclusion of a best guess, rather than seeing narrower ranges in the expert group due to an anchoring effect. Otherwise, we would need to conclude that our experiment adds further evidence to the case *against* anchoring playing any significant role in causing overconfidence. Instead, as has been the case in the majority of instances, we observe that an initial best guess tends to widen rather than narrow subsequently elicited ranges, although by different amounts.

Future Research

As noted above, a key concern with the current analyses relates to the definition of expertise. While the self-ratings that we used did correlate in the expected manner with all of our variables, the fact that a simple 3-point confidence scale was a better predictor is concerning, as is the observation that so few of our sample regarded themselves as being of above average expertise on the task.

To combat this, additional experiments, specifically targeting samples expected to have higher than average knowledge of the domain in question are required, along with pre-experimental testing to directly measure this knowledge. This will enable direct comparisons between people with genuinely high expertise and the general populace and thereby clarify the remaining question of whether true experts will actually be made more overconfident by the inclusion of a best guess in a range elicitation task.

Conclusions

Given the above, it seems reasonable to conclude that expertise does, differentially, affect people's response to different elicitation methods. This is of great importance for the transfer of elicitation techniques between laboratory and applied settings as it suggests that effects observed in the laboratory may not be the same as those seen in practice.

That is, an elicitation effect, shown to be of benefit in laboratory testing, still needs to be tested on experts before we can state, with certainty that it improves elicited values.

[3] Distributions of variables were skewed. Thus, a rank transformation was performed on all observations, with the lowest rank of "1" assigned to the smallest observation.

Acknowledgments

MBW and SHB are supported through ExxonMobil and Santos' funding of the Improved Business Performance group at the Australian School of Petroleum. DJN was supported by ARC grant DP0773794.

References

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49*, 188-207.

Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 928-935.

Clemen, R. L. (2001). Assessing 10-50-90s: a surprise. *Decision Analysis Newsletter, 20*, 2-15.

Conover, W. J. & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximise the accuracy and power of your research. *American Psychologist, 63*, 591-601.

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.

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

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25,* 1038-1052.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110-129.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

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

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

Tetlock, P. E. (1999). Theory-driven reasoning about plausible pasts and probable futures in world politics: are we prisoners of our preconceptions? *American Journal of Political Science, 43*, 335-366.

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

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

Welsh, M. B., Navarro, D. J., & Begg, S. H. (submitted). *Number preference, precision and implicit confidence.*

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

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