

Running Head: NUMBER PREFERENCE, PRECISION AND CONFIDENCE

Number Preference, Data Precision and Implicit Confidence

Matthew B. Welsh, Daniel J. Navarro and Steve H. Begg

University of Adelaide

Australia

Abstract

In elicitation tasks, people make estimates under conditions of uncertainty and their answers are then treated as if they were certain of them. An analysis of patterns of number preferences during elicitation tasks, however, shows a marker indicating how confident the elicitor should be in the elicitee's response. This marker is the precision (number of significant figures) of the estimate. Results of two experiments are presented, which indicate significant relationships between the precision, accuracy and confidence of participant's responses and, further, that precision, as well as being distinct from explicit confidence, can be a better marker of a person's implicit confidence in (and, thus, accuracy of) their answer. The implications of these findings are discussed, in relation to a case study and previous research, for interpreting responses to common decision-making tasks such as tests of overconfidence.

Keywords: number preference, confidence, precision, elicitation, overconfidence.

Number Preference, Data Precision and Implicit Confidence

Imagine that you ask two friends to guess the height of Mt Everest: one says $8\frac{1}{2}$ km, while the other says 9037.21m. Later you check the true answer and discover that Mt Everest is, according to the latest surveys, 8844.43m high. Of your two friends, who would you feel gave the better answer? An obvious way to answer this question would be to focus on accuracy, and see which guess is closer to the true value. Since the first guess is 344.43m too low, and the second guess is 192.78m too high, an accuracy-based evaluation would conclude that the second estimate is superior to the first.

Intuitively, however, something seems wrong here. Clearly, the first person has chosen to give a very imprecise answer whereas the second speaker has given a very precise answer. The first answer appears to be specified to the nearest half-kilometer and thus, in its own terms, is only slightly wrong. Specifically, measured with respect to the person's scale of choice, the error is a single unit: Mt Everest, to the nearest half-kilometer, is 9 km high, not $8\frac{1}{2}$. In contrast, the second answer is specified to the nearest centimeter. In its own terms, the error is thus over 19,000 units. From a pragmatic perspective (Sperber & Wilson, 1986), one might, therefore, feel justified in being less forgiving of the second person's answer. In everyday speech the choice to say "nine thousand and thirty seven point two one meters" rather than the more succinct "nine thousand meters" is highly informative, because it conveys information about the size of the expected error. That is, the listener will assume that the speaker's decision to give the answer to high precision was relevant to the question. Hence the 37.21m difference between 9000 and 9037.21 must be important, which implies that the speaker is highly confident that the true value is very close to 9037m.

Although the above scenario relies on the way verbally-specified answers are given, a similar effect could be observed in written answers. In the decision-making literature, for example, it is known that people prefer to give estimates that are round numbers (Baird, Lewis, & Romer, 1970). A natural assumption, then, is that people might deliberately use the precision (that is, the degree of roundedness) of a written answer to convey information about their confidence in its accuracy. Specifically, it is suggested that both conversational and written estimates may use similar standards of measurement to those used in the physical sciences where, typically, any measurement is given with an error range of +/- half the smallest calibration of the measurement device. For instance, a ruler marked in millimeters is presumed to be accurate to the nearest millimeter +/-0.5mm. This approach to calibration relies heavily on the number of significant digits used: thus, a scientist who specifies that the speed of light is $c = 3 \times 10^8 \text{ ms}^{-1}$ is, implicitly, signaling a much larger potential error than the one who says that $\pi = 3.141592$, and it seems likely that this convention simply mimics similar usage in everyday language.

The main corollary of this phenomenon is that a consideration of the precision of an answer may influence the assessment of its accuracy. Although seemingly unremarkable, this observation has implications for the interpretation of estimates given by participants during a variety of judgment and decision-making tasks. In particular, if observed number preferences (i.e., tendency to use estimates that are multiples of 5, 10, 50, 100, etc; Baird et al., 1970) reflect a participant's desire to convey the level of precision at which their estimate should be considered, then it may not be valid to treat number preference as if it were just a source of random measurement error (as would be the case if people were rounding numbers arbitrarily). Rather, if the precision of an estimate reflects an attempt, by the participant, to communicate the

extent of their knowledge, it should come as no surprise to discover that the “rounding” errors covary systematically with accuracy. As a result, if we simply average across all responses, taking no account of their precision, we may arrive at a misleading picture of peoples’ beliefs and decision making abilities.

In this paper, we present two experiments that illustrate the fact that people do use precision to convey information about accuracy, followed by a reanalysis of a previous data set that shows the extent to which a naïve data analysis – one that ignores participants’ systematic use of imprecise estimates – overestimates the extent of overconfidence in a traditional, interval estimation task (Lichtenstein, Fischhoff, & Phillips, 1982).

Numerical General Knowledge

The primary aim of this, first experiment was to see if people's number preferences are, in part, a reflection of the precision-as-confidence hypothesis. Specifically, when asked to make estimates, does the use of “round” numbers correctly predict the accuracy of the answer and a person's confidence in it? A secondary concern was whether the information yielded by a measure of precision would yield the same or different information than an explicit, self-rating of confidence; and whether the provision of an opportunity to provide such might alter the level of precision at which participants chose to respond (the expectation here being that allowing participants to explicitly state their confidence level might make them more likely to give more precise answers).

Method

Participants

Participants were 36 undergraduate University of Adelaide students, 4 males and 32 females, with a mean age of 22.7 (SD = 4.8). All participants received a \$20 book voucher for their participation in this and another, separate, experiment.

Materials

Two sets of ten questions were written for the experiment, all of which had 4-digit integers as answers. Each question was half of a pair (one from each set) designed to be of approximately equal difficulty (being on the same sub-topic and relating to the same region) to enable the ordering pattern described in the procedure section below. Within each set of questions, four were questions about historical events, four were about geography and two were about popular culture. Questions within each of these topics were divided equally between those about Australian and world events. The selection of questions was deliberately varied so as to attempt to produce a range of levels of difficulty, leading to a wide range of accuracy and confidence in the sample's answers.

Procedure

Participants were randomly assigned to one of two groups, and questions were presented via computer. In the first part of the experiment participants were asked only to make a single estimate in response to the question whereas, in the second, they were also asked to provide confidence ratings, on a scale of 0 (no confidence) to 10 (very confident). Both groups answered all 20 questions: they differed solely in terms of which of the two sets of questions appeared in

the first part of the experiment, and which set appeared in the second. Most participants completed the task within 20 minutes.

Results

Scoring

In total, the experiment involved 720 cases (36 participants x 20 questions) which, for the sake of simplicity, we treat as independent observations rather than fit a complicated hierarchical model (discussed later). For each such case, we constructed three measurements: confidence (available for half the cases), accuracy and precision. The first was simply the explicit *confidence* rating given by participants to indicate how confident they were in the accuracy of their estimates – on a 0-10 scale, with high values corresponding to high confidence. Secondly, the *accuracy* of participants was assessed by calculating the absolute difference between their estimate and the true answer on each question. Low accuracy scores thus indicated more accurate answers. In determining how to judge *precision*, we took the simplest approach and decided that we would record, as an estimate's precision, the number of trailing zeros. That is: estimates that had non-zero final digits (e.g., 7, 132, 1456, etc) were scored as 0 on precision; multiples of 10 (e.g., 30, 110, etc) scored 1; multiples of 100 (e.g., 600, 1300, etc), scored 2; multiples of 1000 (e.g., 4000, 12000, etc), scored 3; and so on. Thus, low values indicate more precise answers. In view of these scoring conventions, we would expect precision and accuracy to be positively related to one another while confidence would be negative related to both.

Overview of the data

One of the most striking characteristics observed in the data is the fact that, where errors were large, the last digit in the number was almost certainly a zero. This is shown in Figure 1, which plots the probabilities for each digit as a function of the accuracy of estimates. The use of

zero increases dramatically as accuracy decreases, while all other last-digit probabilities decrease. Moreover, even when errors are low, the frequency of a last-digit zero is over three times the chance base rate. This suggests that the obvious definition of *precision* as the number of “trailing zeros” seems to have accurately captured the expected relationship between precision and accuracy (that is, we need not consider the role played by other digits). Overall, the precision of the estimates tended to be modest, with median value 1 – that is, the typical estimate was a multiple of 10.

---- Figure 1 about here ----

An additional observation from an initial examination of the data was that, for the most part, participants found the questions quite difficult, with the result that the median absolute deviation from the true answer was 264 and the interquartile range ran from 35 to 2277. Given this variability, accuracy scores were converted to a logarithmic scale in order to allow easy visualization: perfect accuracy was scored 0; an error between 1 and 10 scored 1; between 11 and 100 scored 2; etc. Not surprisingly, when converted to the logarithmically-scaled accuracy scores described above, the median accuracy was 3 (i.e., error in the 101 to 1000 range).

Finally, confidence values, as would be expected given the apparent difficulty participants had with the questions, also tended to be low, with only 37% of confidence scores exceeding 1.

The general pattern of results is shown in Figure 2, which plots the distribution of all precision, confidence and accuracy scores, as well as bubble plots showing the relationship between all pairs of variables. Table 1 confirms that all pairs of variables are related, with significant, weak to medium, Spearman rank-order correlations between all three.

----- Figure 2 about here -----

----- Table 1 about here -----

To determine whether the inclusion of explicit confidence ratings altered the answers given, we conducted Wilcoxon rank sum tests for each question. In terms of the median estimate, not one of the 20 questions showed a significant effect at the $\alpha=.05$ level. That is, the groups of participants who answered a particular question with and without the opportunity to provide a confidence rating were equally accurate.

For the precision of the estimates on the other hand, 3 of the 20 questions showed significant effects, although none were particularly large. Additionally, the direction of these significant effects were all in the opposite direction to that predicted – that is, when allowed to use confidence ratings people were sometimes *less* precise. Given that the questions with confidence ratings were always delivered as the second group, this could be taken as indicating a weak order effect but, given that 5 of the 17 non-significant results were in the initially predicted direction and another 3 showed close to zero difference, this would be a long bow to draw.

Additionally, 3 significant effects from 20 tests at the .05 level is not inconsistent with an overall null effect (i.e., the probability of 3 or more false positives from 20 tests at the .05 level under the null hypothesis is greater than 5%). In short, adding confidence ratings appears to have no effect on the typical magnitude of the estimates given and, most probably, has little to no effect on the precision.

Precision & confidence.

As noted previously, there is a weak relationship ($\rho=.15$) between *confidence* and *precision*, of a similar magnitude as that between *confidence* and *accuracy* ($\rho=.12$). The bubble plot in Figure 2 provides a sense of how the two are related: however, since the plot is somewhat dense, it is helpful to collapse the confidence ratings into the 231 least confident answers (ratings

of 0 or 1) and the remaining 129 answers (ratings ≥ 2). The frequencies of different precision levels for each group are shown in Figure 3.

Looking at this figure, one sees a clear difference in precision by confidence group; a 2x4 χ^2 -test confirms this, $\chi^2(3) = 16.6$, $p < .001$. In cases where participants expressed at least some confidence (≥ 2) in their answers, 44% of the answers were specified to full precision (i.e., did not end with a zero), as opposed to only 26% of the very low-confidence answers. Conversely while in neither case are there very many answers specified only to the nearest 1000 (i.e., precision 3), the proportion of answers given with middling precision (1 or 2) is 60% when confidence is very low, and falls to 41% when there is greater confidence in the answer.

----- Figure 3 about here -----

Precision as mediator

In a key sense, the role of both confidence and precision in an elicitation context is as predictors of accuracy. In applied contexts, accuracy tends to be hard to measure directly (given that, if one knows the true answer, one rarely needs to seek additional opinion). With that in mind, a natural question to ask is whether the two different predictors carry different information about accuracy. The correlations in Table 1 suggest that precision is the better predictor, but some care is required in order to tease out the relationships. To illustrate this, Table 2 shows the partial rank-order correlations between all pairs of variables, while controlling for the third. Note that the relationship between confidence and precision does not change, and neither does the relationship between precision and accuracy. However, when precision is controlled for, the relationship between accuracy and confidence vanishes.

----- Table 2 about here -----

Minimal effects of rounding error

One final possibility that needs to be considered relates to potential artifacts due to rounding error. For instance, suppose we really did think that Mt Everest stands at 8844m in height. If we are confident in our guess, we might report a high precision answer and as a result we would be in error by less than 1m. However, if we are very uncertain and instead report the low precision answer of 9000m, the error jumps to 156m. That is, *underconfidence* could indirectly cause errors due to unnecessarily imprecise response coding. Intuitively this seems unlikely since overconfidence is more common but it is, nevertheless, important to check for this possibility.

----- Figure 4 about here -----

In a straightforward test of the rounding-error hypothesis, we took all participant responses, rounded them to the nearest 1000, and calculated the error of these artificially-imprecise estimates. We then plotted the distribution of errors as a function of the original (i.e., pre-rounding) precision. The results are shown in Figure 4: even though the errors are calculated using data of equivalent precision, there is a strong relationship between the original precision and the new error. In fact, the strength of the rank-order correlation is barely affected by this manipulation ($\rho=.48$, $p<.001$) and, as before, controlling for confidence makes no difference ($\rho=.49$, $p<.001$). Thus, there is no evidence that underconfidence is contributing to the relationship between precision and accuracy.

Within-subject effects

As a final issue, it is worth briefly noting that this experiment is technically a mixed between- and within-subject design, with each of the 36 participants responding to the same set of 20 questions in one of two orders. In this situation, correlations may be induced due to

question effects (some questions are harder), participant effects (some people have better general knowledge than others), or interactions between the two (some people find some questions harder than others). It is important to recognize that – while these are different ways of inducing the effect – these differences are ancillary to the topic at hand. Our main interest is in the correlation between the different measures.

That said, our expectation was that the interesting individual differences would tend to operate at the level of “participant by question” level rather than at the participant level. That is, we expect that the relationship would be observable within participants as well as between them. We confirmed this by calculating rank order correlations comparing individual participants’ accuracy and precision scores. The median correlation was $\rho = 0.45$, $IQR = [0.32\ 0.59]$, and the data from 33 of the 36 participants showed the expected, positive, relationship between accuracy and precision, which a sign test indicates is significant, $p < .001$.

Discussion

The general pattern of results suggests that people's number preferences in an estimation task do not reflect an arbitrary fondness for the decimal system (Albers, 1999). Rather, the preferences are systematic, and encode an implicit confidence rating. What the data appear to suggest is a situation in which precision mediates the relationship between confidence ratings and accuracy. We would explain this pattern in terms of a single latent variable *subjective confidence*, which people use to choose both their response precision and confidence rating. If (as one would expect) subjective confidence reflects metacognitive knowledge about expected response accuracy, then the response precision and the confidence rating would both correlate with the observed accuracy. However, if it is the case that the relationship between *subjective confidence* and chosen *precision* is actually much stronger than that between *subjective*

confidence and explicit confidence ratings, then controlling for precision would eliminate the correlation between accuracy and rated confidence – as was observed here. This strikes us as the most plausible explanation; although it suggests the surprising possibility that in some contexts, at least, intuitive response precision is a better marker of latent confidence than explicit judgment.

Memory for Numerical Knowledge

A concern regarding the results from the first experiment, however, was the limited range of responses for the explicit confidence ratings. By a large margin the two most frequent confidence ratings were 0 or 1, even though the response scale ranged from 0 to 10. As a result, we wondered whether the main finding – that precision carries additional information about accuracy even after controlling for confidence – might be a restriction-of-range artifact. That is, if we could alter the task so that people felt more confident in their answers, would the effect disappear?

We consider this issue in the study reported below. Although we did consider designing a new set of “easier” general knowledge questions, we were skeptical that this would be any more successful than the first experiment. As a result, we adopted a different approach, altering the nature of the task to boost confidence. Specifically, we provided participants with the answers in advance, effectively altering the task to a test of memory. Moreover, having encountered similar problems in other experimental designs, a dataset with the necessary characteristics (a wide range of confidence ratings) was already available, from an unpublished experiment on the anchoring effect. We describe the relevant aspects to the experiment here (i.e., the control condition), and analyze the data from a “precision and confidence” perspective. The appendix contains additional details relevant only to the anchoring manipulation.

Method

Participants

Participants were 4 graduate psychology students and 11 university-educated members of the general public. These included 5 males and 10 females, with a mean age of 31.5 ($SD = 7.4$), each paid \$20 for participation.

Materials

As with the first experiment, the task involved asking people general knowledge questions requiring numerical responses, and asking for confidence ratings in addition. Once again, the questions covered a range of topics (geography, science, sports trivia and history), though the true answers now involved a wider range of numbers, from single digit to four digit answers with some also including decimal points. In total there were 54 such questions.

Procedure

To counter the tendency of participants to give very low confidence ratings on general knowledge tasks, participants in this task were shown all 54 general knowledge statements, including the answers, at the start of the experiment. After this initial study phase, participants were presented all 54 questions and asked to give their best estimate of the answer, along with a confidence rating (from 0 to 10). Thus, rather than relying on general knowledge, the task became a memory test. This was done specifically to increase the frequency of confident and accurate answers. The 54 questions were divided into three blocks of 18 questions, two of which were treatment conditions involving an anchoring manipulation (not relevant to the current analysis: see appendix) plus a control condition. The order of the three blocks was

counterbalanced. Only questions from the control conditions were included in the following analyses – that is, 18 per participant.

Data cleaning

This procedure yielded 270 data points (18 responses x 15 participants). However, careful examination of the data led to the removal of 9 data points in which participants gave “unreasonable” responses. The unreasonable responses included several estimates of “0”, a small number of responses that took the form 2345 or 5678 (adjacent keys on the keyboard) when these were massively inconsistent with the true answer and two entries that were clearly impossible (a date of 4356AD and a percentage of 188%). All 261 other responses were included.

Results

Scoring

Accuracy and confidence scores were calculated exactly as described in experiment 1. That is, confidence was simply the value from 0-10 given by the participant, while *accuracy* was the absolute difference between a participant’s estimate and the true answer to a given question.

For precision, however, the story is a little more complex. The *observed precision* values were calculated in the same way as for study 1, by counting the number of terminal zeros. However, due to the inclusion of questions that had decimals in their answers, the precision values were allowed to be negative. That is, an observed precision score of -1 indicated an answer that was given to one decimal place, and -2 indicates that the answer was given to two decimal places. However, given that the *expected* answer now also varied in precision (-2, -1 or 0), we also considered the *relative precision*, namely the difference between the precision of the correct answer and the one given by participants. Thus, if the correct answer given to participants

at study was 36.36%, and an answer of 39.3% was provided, then the *expected* precision is -2, the *observed* precision is -1, and the *relative* precision is 1. Finally, for use in assessing linear relationships, *absolute* precision was calculated. This was simply the modulus of relative precision and was used to ensure that answers that were more precise than the true answer were not regarded as beneficial.

As can be seen in Table 3 over half of answers were appropriately precise and, moreover, there are not large differences between the measures.

----- Table 3 about here -----

Within subjects effects

As with the previous study, the main interest is the question of whether there is a correlation between precision, accuracy and confidence across the 261 valid responses. The question of whether this is caused by a consistent difference among questions (some questions might be inherently harder) or participants (some people have better knowledge), or a collection of random interactions (some people find some questions harder than others) is of secondary interest, and not one that we consider here. Thus, as was the case in Experiment 1, the dataset was treated as if it consisted of 261 independent observations.

Correlations among precision, confidence and accuracy

We begin by considering the accuracy-confidence relationship (Figure 5) and the accuracy-precision relationship (Figure 6) separately. Looking at Figure 5, one sees the expected relationship between accuracy and confidence – with the magnitude of errors (i.e., the accuracy score) generally decreasing where participants were more confident in those answers. Similarly, in Figure 6, the size of the errors made by people is clearly related to the absolute precision with

which they answered the questions, with more precise answers tending to be the most accurate (recalling, of course, that low numbers indicate high precision).

---- Figure 5 about here ----

---- Figure 6 about here ----

Table 4 quantifies the relationships between the three variables, in terms of rank-order correlation: all three variables are significantly correlated with one another in the expected direction – that is, when giving more precise answers, people also make smaller errors and are more confident in their answers. As with experiment 1, a key question is whether this is one effect or two – that is, whether confidence and precision convey the same information about accuracy, or whether there is distinct information in both. To test this, the correlations between accuracy and confidence and accuracy and precision were recalculated, controlling in each case for the other. The partial correlation between accuracy and precision was at $\rho = 0.25$ ($p < .001$) once the effect of confidence was controlled for. Similarly, the relationship between confidence and accuracy remained at $\rho = 0.30$ ($p < .001$).

---- Table 4 about here ----

Given these results, it seemed sensible to examine all three variables simultaneously. To do so, both precision and confidence were sorted into groups representing low, medium and high responses according to simple rules of thumb. Specifically, confidence values of 0-3 were assigned to the *low* confidence group, 4-6 *medium* and 7-10 *high*. For precision, relative scores of -1 or 0 were assigned to the *high* precision group, 1 to *medium* and 2-3 to *low*. Figure 7 thus shows the mean errors achieved by each precision-by-confidence group.

---- Figure 7 about here ----

Looking at Figure 7, one generally sees the expected relationships; that is, within each precision group, one sees that, as confidence falls, errors are observed to rise. Similarly, within each confidence group, people tend to be more accurate where they are more precise. There is, however, one exception to this trend – the high precision-low confidence group. This group is markedly less accurate than one would expect from looking at the remainder of the figure. That is, there seem to have been a significant number of occasions when participants have given very precise answers that they, rightly as it turns out, had little confidence were accurate.

To assess these relationships, a 3x3 ANOVA was run, comparing the mean absolute errors of participants in each of the groups. This indicated that the main effects of both confidence and precision were significant, $F(2, 252) = 7.95$ and 3.44 , $p < .001$ and $p = .006^1$, but the interaction between them was not, $F(4, 252) = 1.18$, $p = .321$.

Discussion

The results of study 2 make clear that precision and confidence, while related to one another, are not exactly the same thing. Whether this is because one is more explicit and conscious application of metacognition while the other is an unconscious adjustment remains an open question.

Given a task where the full range of confidence ratings is available, confidence proved to be a slightly better predictor of accuracy than precision, with a correlation of 0.37 compared to 0.34 but, even taking into account the two variables shared variance, precision remains a significant predictor of accuracy – a finding supported by the observation from the 3x3 ANOVA that the main effects of confidence and precision are both significant. That said, study 2's results replicate the observation from study 1 that, given a restricted range of confidence judgments,

precision may prove a better predictor of accuracy. That is, these findings are likely to be of added importance when confidence is uniformly low.

While the interaction between confidence and precision is not significant in this dataset, the strange peak in Figure 7 where high precision and low confidence are linked to inaccurate responses is still of sufficient interest to require further discussion. A likely cause of this is ‘random’ responding in situations where the participant had no idea of the true answer. Recall that, during data cleaning, a number of responses were removed which had the form 2345 or 4567 – that is, runs of numbers across the keyboard. These would have been coded as high precision responses but were, instead, regarded as effectively random responses by participants and disregarded. It therefore seems possible that similarly random answers that did not produce such an obvious a pattern may have caused the peak under discussion.

That is, the peak is best explained as a proportion of responses conforming to an alternate strategy for answering when ‘stumped’ by the question. Some participants, as we hypothesized, resorted to rounded numbers indicating the imprecision of their estimates whereas others may have given essentially random responses and indicated this using their confidence rating. An interesting possibility is that this may reflect a difference in conscientiousness regarding the task at hand. That is, less conscientious people may be tempted to give random responses and use the confidence rating to indicate this while more conscientious people attempt to give as good an answer as they can under the circumstances.

Number Preference & Overconfidence

The elicitation of uncertainty describes the process of converting a person's subjective beliefs regarding uncertain events into a numerical form, generally with the goal of incorporating these beliefs into quantitative statistical analysis (Wolfson, 2001). Elicitation methods are often

used to assist probabilistic forecasting in fields such as petroleum exploration (Attanasi & Schuenemeyer, 2002) and meteorology (Morgan & Keith, 1995). There are a number of different methods available but the technique most commonly used in the oil and gas industry, for example, is a simple elicitation of 80% confidence ranges (Capen, 1976; Hawkins, Coopersmith, & Cunningham, 2002). In this method, the elicitee is asked to give values x and y , such that they are 80% certain that the quantity of interest lies between x and y .

The major problem with such methods is that they tend to produce overconfidence (Lichtenstein, et al., 1982). That is, people generally give ranges that are too narrow, such that true values tend to fall outside the range much more frequently than the expected 20% rate. As a result, much of the literature on elicitation focuses on mechanisms for overcoming uncertainty or “debiasing” participants. Various techniques from simple exhortations to widen ranges (Lichtenstein, et al., 1982) through repeated feedback (Murphy & Winkler, 1977) to the use of probabilistic games (Hawkins, et al., 2002) are recommended. In general, however, such techniques reduce overconfidence but do not eliminate it (Morgan & Henrion, 1990).

The causes of overconfidence are still debated (for a recent discussion, see, e.g., Heywood-Smith, Welsh, & Begg, 2008) but, if people use imprecise numbers to reflect their own uncertainty regarding the estimates they make, as demonstrated above, this would contribute to apparent overconfidence. To understand why this is, consider, for instance, a person who has given an 80% confidence interval of 100-500. Statistically, it is extremely implausible that both endpoints of a precise range are multiples of 100, yet people often give these sorts of numbers.

The implication of this for an elicited range, of course, is that the person who gives a 100-500 range and states that they are 80% certain that a measured value will fall within that range may not, in fact, be 80% confident that the value will fall within that *precise* range. Rather,

following our discussion above regarding the standard conversational and scientific rules of measurement, it may be more reasonable to conclude that they are 80% confident that the value will fall with the range 100 ± 50 to 500 ± 50 ; that is, in the 50-550 range. Treating imprecise range endpoints as precise would thus lead to apparent overconfidence as some proportion of values that the elicitee implicitly considers part of their range are excluded at either end.

The results obtained here may also help to explain an interesting observation from the overconfidence literature: the impact of format dependence (Juslin, Wennerholm, & Olsson, 1999). That is, observations such as: that the same participant will give a different answer when asked to generate a confidence interval than when asked to evaluate that same confidence interval (Winman, Hansson, & Juslin, 2004). For example, a person asked to generate an 80% confidence range may state that they are 80% confident that a value will fall within the range 150 to 350. If later asked, however, how likely the value is to fall within that same 150 to 350 range, they are likely to be less than 80% confident.

A possible explanation for this effect was provided by Winman et al (2004), who argued it results from the statistical naivety of participants; and simulation of the effect (on overconfidence) of using sample variances to estimate population variance can be seen in Welsh, Begg, Bratvold and Lee (2004), where it is shown that sample sizes in the range of human short-term memory limitations do lead to overconfident estimates of population dispersion. This explanation, however, requires that people think in peculiarly statistical ways. Specifically, for overconfidence to be the result of sampling from memory, assumptions must be made about the nature of memory and recall that do not necessarily accord with mnemonic theory and experimental results (for a discussion of this, see, e.g., Bruza, Welsh, & Navarro, 2008).

An alternate explanation can also be made, however, invoking the concept of data precision described above. If the numbers *generated* by people incorporate precision as markers of accuracy but, in contrast, people interpret the numbers to be *evaluated* by them in an evaluation task as being more precise (as could be the case given the role assigned to a participant by such a task), then this would lead to a disparity between the levels of overconfidence observed in generated and evaluated ranges, as has been observed.

To demonstrate the potential of such a discrepancy to affect measurement of the calibration of an individual or group, we conducted a case study, using data from a previous paper on overconfidence to see what impact including the precision of answers would have on the calculation of calibration.

Method

Dataset

The data used for this analysis was taken from Welsh, Bratvold and Begg (2005) where a sample of 123 oil and gas industry workers were given a 10-question overconfidence questionnaire as part of a larger survey on cognitive biases affecting petroleum industry decision making. The sample consisted of 18 females and 105 males, with a mean age of 39.8 ($SD = 9.4$) and an average of 13.5 years of industry experience ($SD = 9.1$).

The values asked for in the overconfidence questions all related to the petroleum industry and all of these questions asked participants to give 80% confidence intervals - that is, they were asked to provide a high and a low value such that they were 80% confident that the correct answer to the question would lie within the indicated range.

Recoding

The calibration data from the previous study was examined and each point estimate (the high and low end-points of each range) was examined to determine its degree of precision. This was done as defined in the second study herein, by describing precision as the number of trailing zeros on an estimate. Then, for each case where the precision score was 1 or higher (indicating that the estimate was a multiple of 10, 100, 1000, etc), the estimate was adjusted to include additional points within one half of 10 to the power of the precision magnitude.

For example, a low-end estimate of 300 would be scored as having a precision of 2 and then adjusted by *subtracting* one-half of 10^2 - that is, 50 - resulting in a new low-end estimate of 250. A high end estimate of 1000 (precision of 3), by comparison, would have a one-half of 10^3 , that is 500, *added* to it resulting in a new high-end estimate of 1500. The new high and low end estimates were then used to recalculate participants' calibration across the set of 10 questions for comparison with the original calibration scores.

Results

The calibration scores achieved by participants on the 10 questions, both in terms of their raw estimates and those adjusted to take into account differences in precision, were compared. Figure 8 shows the mean calibration of the participants (the proportion of participants' '80%' ranges that actually contained the true value) both before and after adjustments for precision.

Looking at this figure, one sees that - as would be expected - the widening of the ranges resulting from the precision adjustment has improved the calibration on each question. Of course, the precision adjustment described could only ever widen ranges (or leave them unchanged) so the question of interest is: how much difference does this adjustment make to calibration?

The degree of improvement varied from 3.4% to 16.5% across the ten questions, with a mean improvement of 9.0%. To determine the significance of these changes, binomial tests were conducted to calculate how often one would expect to see as many correct answers as were achieved using the precision adjustment, given the base calibration rate in the raw data. Table 5 shows these results - indicating that in eight out of the ten questions, the difference between the calibration in the adjusted and unadjusted data is significant at the .05 level. The tests also approach significance ($p < .1$) for both of the remaining questions.

---- Figure 8 about here ----

---- Table 5 about here ----

Discussion

Given the above results, it seems clear that the incorporation of precision into calculations of calibration would have a significant effect in terms of reducing overconfidence. A 9% reduction in overconfidence was observed across the 10 questions – with an improvement of up to 16.5% on specific questions. Such an adjustment accounts for the majority of the differences in overconfidence between evaluated and generated ranges described by Winman et al (2004) in their first experiment – which (judging from their Figure 3) range from ~5 to 18%.

That is, assuming that participants do, as suggested, treat an evaluation task as involving precise numbers but naturally tend to generate imprecise ranges, this would account for the majority of the difference in calibration between these two situations.

General Discussion

The results of the above studies seem clearly to support the hypothesis that people tend to use imprecise numbers when attempting (consciously or unconsciously) to convey a lack of confidence in the accuracy of their estimate. Perhaps the most interesting observation was that

this preferential use of round numbers proved a better indicator of the accuracy of a response than an explicit confidence rating (even after controlling for any impact of rounding) where confidence ratings were restricted. That is, in situations where people are uniformly confident or doubtful of their ability to answer correctly, the precision of their answers still offers a mechanism for determining how accurate their answer is likely to be.

Even where people's explicit confidence ratings vary sufficiently for these to be used to predict accuracy, however, the precision of any estimates still adds to our understanding of an individual's belief in their answers' accuracy, indicating that precision and explicit confidence ratings measure somewhat distinct aspects of what we earlier called *subjective confidence*. Thus, even where confidence ratings have been recorded, the precision of estimates still needs to be considered.

Of additional interest is the observation that a reinterpretation of calibration scores incorporating adjustments for the level of precision in participant's estimates significantly improves calibration and could, as was argued above, account for the greatest part of the differences between observed levels of overconfidence in range-evaluation and range-generation tasks - as defined by Winman et al (2004). That is, these results suggest that part of the apparent overconfidence observed in elicitation tasks is, in fact, resulting from the *elicitor's* misunderstanding of exactly how the *elicitee* intends that their answers be interpreted. As such, simply understanding that people's estimates contain markers of accuracy will allow elicitors to better understand the true range of possible outcomes represented by an elicited range.

In terms of future directions, it may be possible to extend this work to look at the implications of number preferences for accuracy and confidence at a finer-grained level. Specifically, while the effects shown herein are clear, the definition of precision as the number of

trailing zeros on an estimate ignores the observations from previous analyses of number preference (Plug, 1977) that multiples of 5 can also be disproportionately represented amongst responses – most strongly in elicitation tasks utilizing 0-100 scales. Inclusion of finer gradations of number preferences may, therefore, further improve our understanding of how accurate estimates should be regarded as.

There is also the possibility that number preferences of the type shown here can explain part, at least, of the confusing relationship between anchoring and overconfidence. That is, the observation that asking for a best guess first sometimes reduces overconfidence (see, e.g., Block & Harper, 1991; Welsh, et al., 2004) but, in other cases, increases it (see, e.g., Heywood-Smith, et al., 2008; Russo & Schoemaker, 1992). To understand how these findings relate to this effect, one must understand that, to the extent that a person's estimates are imprecise, this limits the *minimum* width of their elicited ranges. For example, a person who rounds their estimates to the nearest 100 to represent their uncertainty about the true values might give a range of 100-200. If asked for a best guess first, by comparison, they will be forced to select (at the same level of precision) *either* 100 or 200 and then, when asked to provide a range be forced to give a wider range of 0-200 or 100-300 (depending on whether they selected 100 or 200 as their best guess). Such an effect could override any impact that the best guess has as an anchor on the end-points. Thus, the prediction would be that, where people show preferences for round numbers they are more likely to show reduced overconfidence when asked for a best guess first, whereas people tending to give more precise answers may show increased overconfidence due to the best guess acting as an anchor.

In conclusion, the results presented above lead us to believe that the precision with which people specify their answers does act as a marker of a person's underlying confidence in their

answer's accuracy and is, as such, a good signal regarding the accuracy of the answer - even after controlling for possible artifacts of rounding error and the separate effect of explicit confidence ratings. It seems that, when people are asked to provide numerical estimates, their answers can encode two different things; the estimate itself, and a degree of trust in that estimate. Treating an answer of this form as if it were a "simple" estimate that can be averaged across people or questions is, thus, almost certainly a bad idea and improvements in elicitation seem likely to result from a principled incorporation of the relationships between precision, accuracy and confidence described above.

References

- Albers, W. (1999). Prominence theory as a tool to model boundedly rational decisions. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: the adaptive toolbox*. Cambridge, MA: MIT Press.
- Attanasi, E. D., & Schuenemeyer, J. H. (2002). Some aspects of resource uncertainty and their economic consequences in assessment of the 1002 area of the Arctic National Wildlife Refuge. *Natural Resources Research*, 11(2), 109-120.
- Baird, J. C., Lewis, C., & Romer, D. (1970). Relative frequencies of numerical responses in ratio estimation. *Perception and Psychophysics*, 6, 78-80.
- Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49, 188-207.
- Bruza, B., Welsh, M. B., & Navarro, D. J. (2008). Does memory mediate susceptibility to cognitive biases? Implications of decision-by-sampling theory. In V. Sloutsky, B. Love & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1498-1503). Austin, TX: Cognitive Science Society.
- Capen, E. C. (1976). The difficulty of assessing uncertainty. *Journal of Petroleum Technology*(August), 843-850.
- 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., 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.
- Howell, D. (2004). *Statistical methods for psychology* (4th ed.). Belmont, CA: Thompson Wadsworth.
- Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 1038-1052.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Morgan, M. G., & Henrion, M. (1990). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.
- Morgan, M. G., & Keith, D. W. (1995). Subjective judgements by climate experts. *Environmental Science and Technology*, 29(10), 468A-476A.
- Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26(1), 41-47.
- Plug, C. (1977). Number preferences in ratio estimation and constant-sum scaling. *American Journal of Psychology*, 90(4), 699-704.
- Russo, E. J., & Schoemaker, P. J. H. (1992). Managing Overconfidence. *Sloan Management Review*, 33, 7-17.
- Sperber, D., & Wilson, D. (1986). *Relevance: communication and cognition*. Oxford: Blackwell.

- Welsh, M. B., Begg, S. H., Bratvold, R. B., & Lee, M. D. (2004). SPE 90338: Problems with the elicitation of uncertainty. *Proceedings of the Society of Petroleum Engineers 80th Annual Technical Conference and Exhibition, Houston, Texas: SPE.*
- Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005). SPE 96423 - Cognitive biases in the petroleum industry: impact and remediation. *Proceedings of the Society of Petroleum Engineers 81st Annual Technical Conference and Exhibition.*
- Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: how to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(6), 1167-1175.
- Wolfson, L. J. (2001). Elicitation of probabilities and probability distributions. In E. Science (Ed.), *International Encyclopedia of the Social Sciences* (pp. 4413-4417): Elsevier Science.

Author Note

Matthew B. Welsh, Australian School of Petroleum, University of Adelaide, Adelaide, South Australia 5005, Australia; Daniel J. Navarro, School of Psychology, University of Adelaide, Adelaide, South Australia 5005, Australia; Steve H. Begg, Australian School of Petroleum, University of Adelaide, Adelaide, South Australia 5005, Australia.

MBW and SHB were supported by Santos and ExxonMobil through their support of the Improved Business Performance Group at the Australian School of Petroleum. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). The authors wish to thank Xiaohui Carolyn Chen for her assistance in data collection and Nancy Briggs for her comments on the manuscript.

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

Footnotes

¹ The p-values presented here have been divided by 6 to reflect the fact that the both hypotheses (relating precision and confidence to accuracy) are directional whereas the ANOVA results are not – that is, there is only a 1 in 6 probability that a significant effect would, if due to chance, reveal the predicted order. For justification, see Howell's (2004) *Statistical Methods for Psychology* 4th Edition, page 155. Note, however, that even without this adjustment, both main effects would be significant.

Appendix: Details of the anchoring experiment

Although we omitted the majority of the details of the anchoring experiment used to supply the data for study two (since most of the details are irrelevant) from the main text, we provide those details here for the sake of completeness. The experiment tested 4 graduate psychology students and 11 university-educated members of the general public. These included 5 males and 10 females, with a mean age of 31.5 ($SD = 7.4$), each paid \$20 for participation. To counter the tendency of participants to give very low confidence ratings on general knowledge tasks, participants in this task were, prior to being asked to answer any questions, shown fifty-four general knowledge statements, regarding geography, science, sports trivia and history, presented in a random order using a GUI. These statements all contained numbers, ranging from single digit to 4 digit answers with some also including decimal points. That is, rather than relying on their general knowledge, the task became a memory test where the participants had recently seen all of the answers to the 54 questions that they were subsequently asked to answer. This was done specifically to increase the frequency of high confidence responses. Within the anchoring task, participants saw 18 questions under each of three “knowledge of anchoring” conditions – no knowledge, aware and trained. Within each set of 18 questions, 6 were control questions (with no anchor), 6 had high anchors and 6 low. The order of these sets of 6 questions was varied according to a Latin Square design such that equal numbers of participants saw each type of question (control, high and low) first within each block of 18 questions constituting a condition. After receiving training on all 54 items, participants were shown the questions in the order described above and asked to provide their best guess as to the true answer (after answering a greater/less than question where an anchor was presented) and, additionally, rate their confidence in the accuracy of their answer.

Table 1. Rank-order correlations between precision, confidence and accuracy.

	Precision	Confidence	Accuracy
Precision	-	-.15	.51
Confidence	.004	-	-.12
Accuracy	<.001	.030	-

Note: Upper triangular cells show the correlation ρ , while the lower cells show the p-value. N=

720 for the precision/accuracy case and 360 for the other two cases.

Table 2. Partial rank-order correlations between precision, confidence and accuracy, controlling for the third variable in each case.

	Precision	Confidence	Accuracy
Precision	-	-.11	.53
Confidence	.004	-	-.04
Accuracy	<.001	.460	-

Note: Upper triangular cells show the correlation ρ , while the lower cells show the p-value. N=

360 for all cases.

Table 3. Number of observations by precision score.

Precision	-2	-1	0	1	2	3
Observed	6	7	149	60	27	12
Expected	18	5	187	50	0	0
Relative	0	23	151	40	34	13
Absolute	0	0	151	63	34	13

Table 4. Rank order correlations.

	Confidence	Accuracy	Precision
Confidence	1		
Accuracy	-0.37*	1	
Precision	-0.30*	0.34*	1

Note: N = 261. * - all correlations significant at the .001 level. The

accuracy and precision variables here are both absolute scores.

Table 5. Binomial tests of impact of precision adjustment on calibration by question.

Note: In each case $p = (1 - \text{binomial}(\text{Hits}, N \mid \text{Rate}))/2$, to give the one-tailed probability. Rate is calculated from the Hits and N values in the unadjusted data. N varies between questions because instances where participants failed to give a range were excluded on a case by case basis.

Figure Captions

Figure 1. Evidence for the deliberate use of imprecise estimates. Probability that the last digit in the answer is 0, 1, ..., 9 as a function of the size of the error in the estimate. Error bars denote 95% confidence intervals assuming a beta-binomial distribution. NB – for clarity of presentation, the y-axis of the “0” case differs from that used in the other cases.

Figure 2. Histograms showing the empirical distributions for the precision, confidence and accuracy measures, along with bubble plots showing the joint distributions for all pairs of variables. Bubble diameter reflects number of observations at each point.

Figure 3. The distribution of precision values for the least confident answers (left; confidence 0 or 1) versus the corresponding distribution for the answers in which people expressed at least some confidence (right; confidence ≥ 2). The main difference between the two groups lies in the use of intermediate levels of precision (i.e., either one or two trailing zeros).

Figure 4. Histograms of precision by accuracy controlling for rounding errors. If we convert all estimates to low precision (round every answer to the nearest 1000), and then plot the error of the rounded answers as a function of the precision of the original answer, we still have a strong effect. The leftmost panel shows answers that were originally given in precision -3 (to the nearest 1000), whereas the rightmost panel shows answers originally in precision 0 (nearest 1).

Figure 5. Central measures of accuracy, with inter-quartile range, by confidence rating.

Figure 6. Central measures of accuracy, with inter-quartile range, by absolute precision.

Figure 7. Absolute accuracy of estimates by precision and confidence groups.

Figure 8. Scatter-plot comparing raw calibration scores with calibration scores adjusted for precision on each of the 10 overconfidence questions. Points lying above the line indicate improvements in calibration following the precision adjustment.

Figure 1.

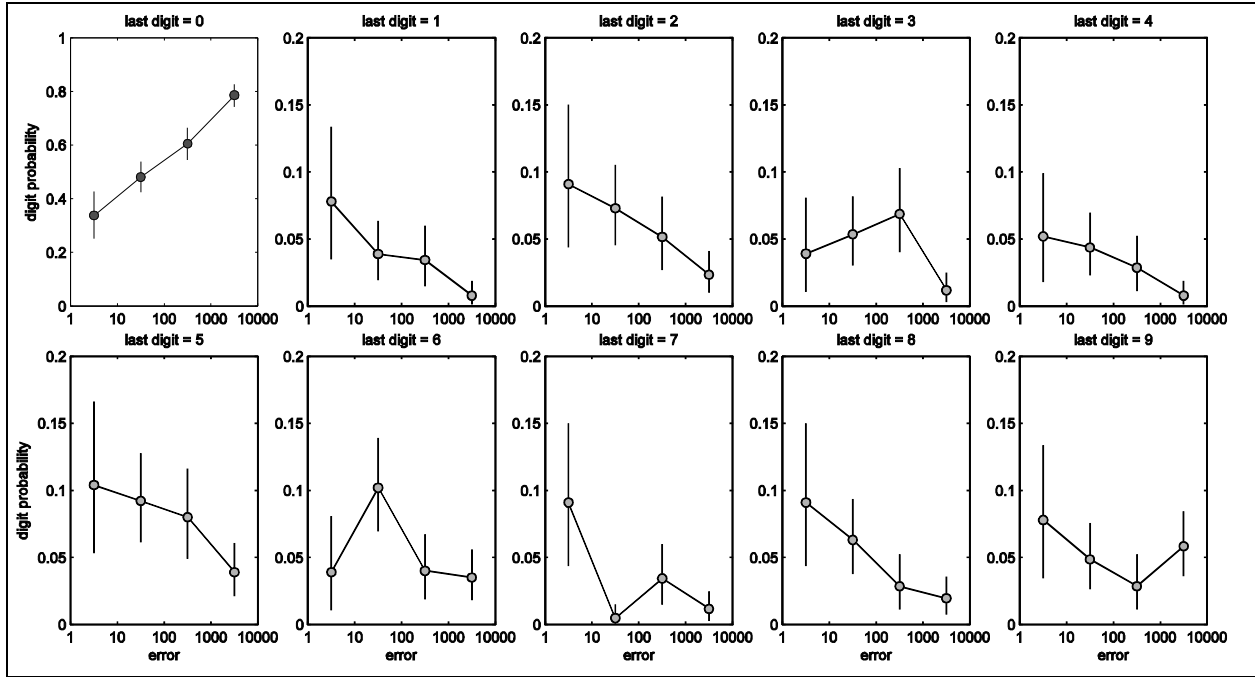


Figure 2.

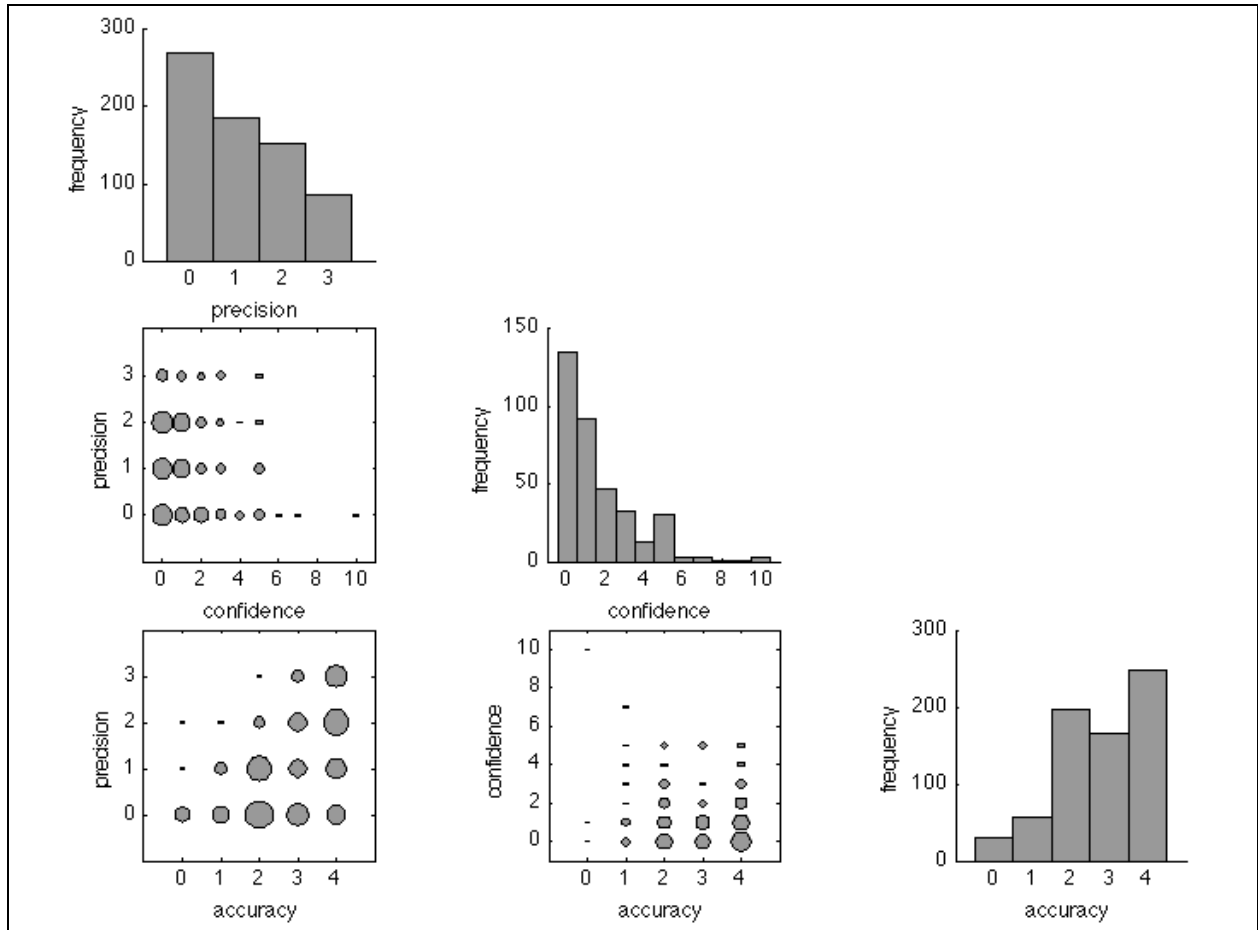


Figure 3.

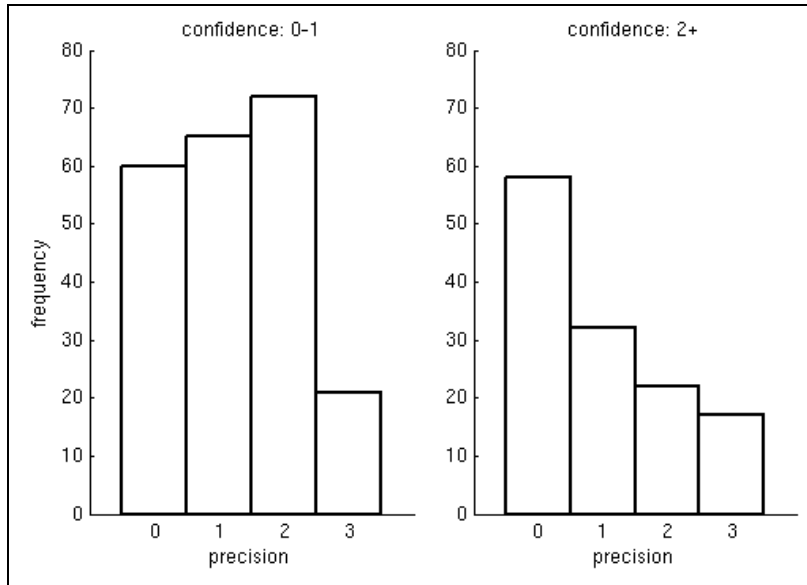


Figure 4.

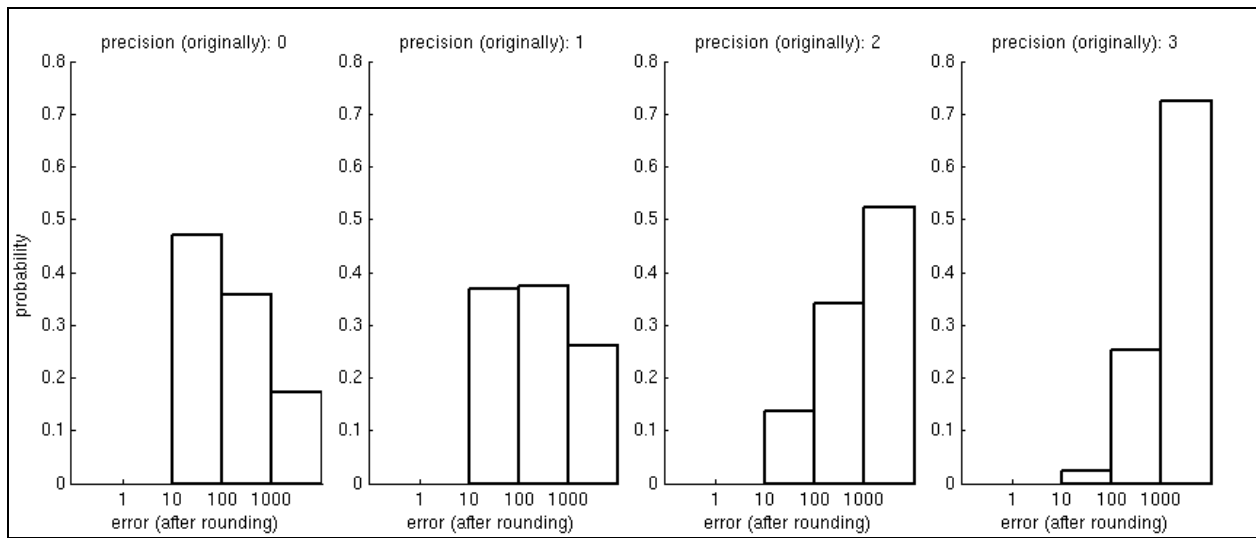


Figure 5.

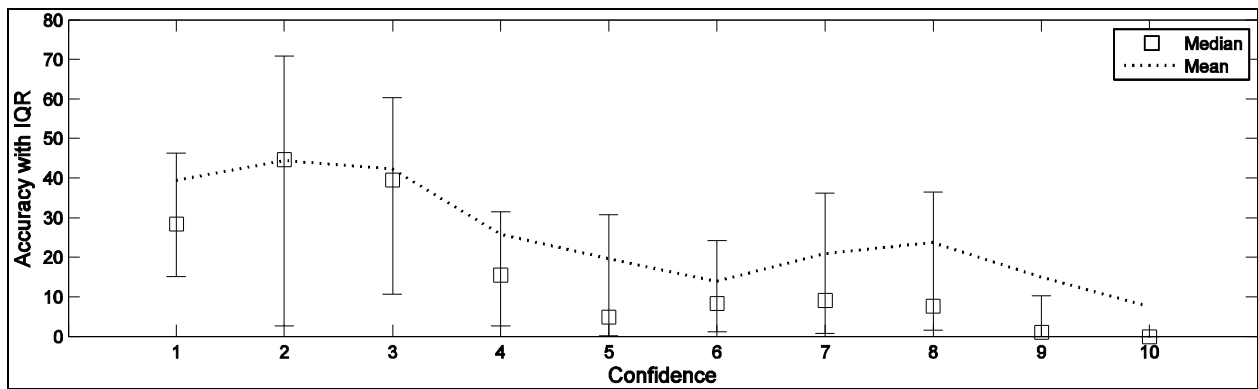


Figure 6.

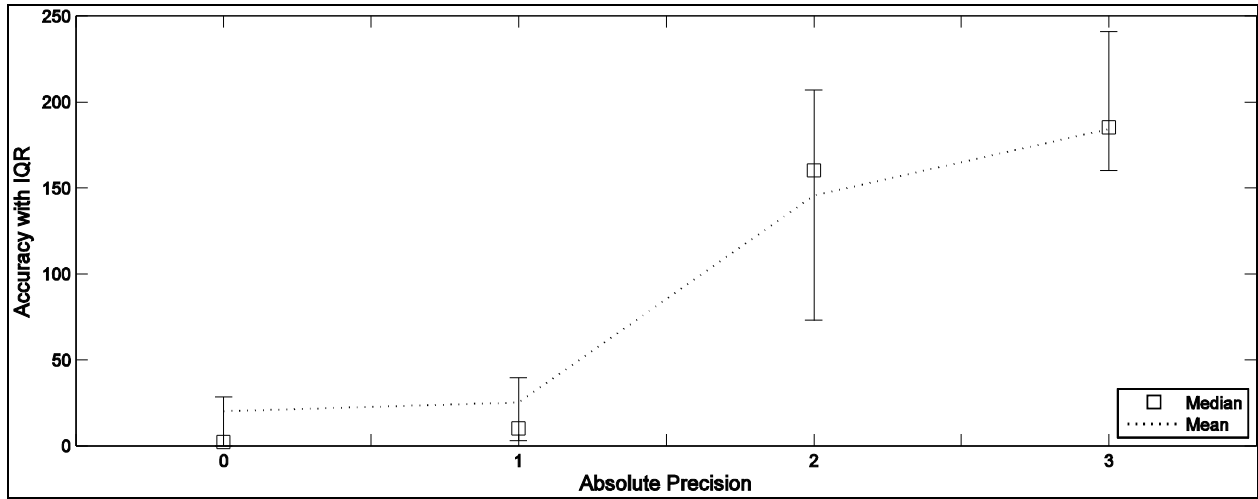


Figure 7.

Figure 8.

