Risk Analysis, Vol. 30, No. 3, 2010 DOI: 10.1111/j.1539-6924.2009.01337.x

Reducing Overconfidence in the Interval Judgments of Experts

Andrew Speirs-Bridge,[1,2] Fiona Fidler,[1,2] Marissa McBride,[2] Louisa Flander,[3] Geoff Cumming,[1] and Mark Burgman[2,*]

Elicitation of expert opinion is important for risk analysis when only limited data are available. Expert opinion is often elicited in the form of subjective confidence intervals; however, these are prone to substantial overconfidence. We investigated the influence of elicitation question format, in particular the number of steps in the elicitation procedure. In a 3-point elicitation procedure, an expert is asked for a lower limit, upper limit, and best guess, the two limits creating an interval of some assigned confidence level (e.g., 80%). In our 4-step interval elicitation procedure, experts were also asked for a realistic lower limit, upper limit, and best guess, but no confidence level was assigned; the fourth step was to rate their anticipated confidence in the interval produced. In our three studies, experts made interval predictions of rates of infectious diseases (Study 1, $n = 21$ and Study 2, $n = 24$: epidemiologists and public health experts), or marine invertebrate populations (Study 3, $n = 34$: ecologists and biologists). We combined the results from our studies using meta-analysis, which found average overconfidence of 11.9%, 95% CI [3.5, 20.3] (a hit rate of 68.1% for 80% intervals)—a substantial decrease in overconfidence compared with previous studies. Studies 2 and 3 suggest that the 4-step procedure is more likely to reduce overconfidence than the 3-point procedure (Cohen's $d = 0.61$, [0.04, 1.18]).

KEY WORDS: Expert elicitation; interval judgment; overconfidence

1. INTRODUCTION

Expert elicitation forms a key step in risk analysis, although elicited intervals often show substantial overconfidence. The critical role of question format in the elicitation of judgments is well supported by the research literature.(1–3) It is possible to achieve substantial reductions in overconfidence by refining question format, with Soll and Klayman(2) and Teigen and Jorgensen(3) producing some of the most impressive reductions to date. Soll and Klayman reduced overconfidence when compared to the "range" question format by 27% using a 3-point interval elicitation procedure; and Teigen and Jorgensen reduced it by 52% using "anticipated" rather than "assigned" confidence levels. See Table I for examples of the question formats. To date, however, these reductions have been observed only in student samples for general knowledge questions. Our studies investigate: (1) whether these impressive reductions in overconfidence can be replicated in expert samples and (2) if further reductions in overconfidence can be achieved by combining the methods of Soll and Klayman with those of Teigen and Jorgensen.

[1]School of Psychological Science, La Trobe University, Australia.
[2]Australian Centre for Excellence in Risk Analysis (ACERA), University of Melbourne, Australia.
[3]Melbourne School of Population Health, University of Melbourne, Australia.
[*]Address correspondence to Mark Burgman, Australian Centre for Excellence in Risk Analysis (ACERA), University of Melbourne, Vic. 3010, Australia; tel: +61-3-8344-7151; markab@unimelb.edu.au.

0272-4332/10/0100-0512$22.00/1

Table I. Overconfidence, Expressed as Confidence Level Minus Hit Rate

Elicitation Format with Examples	Average Overconfidence
Soll and Klayman[2]	
Range (one-point)	41%
I am 80% sure that this happened between ____ and ____.	
Two-point	23%
I am 90% sure that this happened after ____.	
I am 90% sure that this happened before ____.	
3-point	14%
I am 90% sure that this happened after ____.	
I am 90% sure that this happened before ____.	
I think its equally likely that this happened after or before ____.	
Teigen and Jørgensen[3]	
Assigned 90% confidence level	67%
The year Einstein was born is between ____ and ____, with 90% certainty.	
Assigned 50% confidence level	27%
The year Einstein was born is between ____ and ____, with 50% certainty.	
Participant assigned confidence level	15%
The year Einstein was born is between ____ and ____.	
I am ____% confident that my interval contains the true answer.	

Interval judgments are useful because they contain information about uncertainty that is not provided by point estimate judgments, but are subjectively easier than full distributions for experts to produce. They are commonly used as part of procedures for eliciting expert opinion in risk assessment settings. For example, Morgan and Keith[4] in their elicitation of climate experts asked participants for their best estimates with 90% confidence intervals. However, interval judgments are highly susceptible to overconfidence. As Soll and Klayman[2] concluded, there is "little doubt that overconfidence predominates in interval judgments. For example, judges' 90% intervals typically contain the correct answer less than 50% of the time." (p. 299). We wanted the experts' intervals to accurately reflect the full extent of the external and their internal uncertainty about the parameter in question. The goal of our studies was, therefore, to develop an elicitation protocol to minimize overconfidence in our experts' intervals, as measured using hit-and-miss calibration.[4] By "hit-and-miss calibration" we mean that if an expert provided 10 80% confidence intervals, 8 of those 10 should contain the true values for the expert to be perfectly calibrated. If instead the expert's intervals only contained the true values 6 of 10 times then we consider them 20% *overconfident*.

2. OVERCONFIDENCE: A BRIEF OVERVIEW

There is a large literature, dating back several decades, on overconfidence, largely devoted to identifying its moderators. For example, overconfidence increases with the availability of information;[7,8] increases as questions become more difficult;[9] is greater in the absence of regular systematic feedback;[9–11] and can be influenced by the experts' *cognitive style*.[12] Although this list is by no means exhaustive it does highlight the many moderators of overconfidence, and their potential adverse impact on the elicitation process. More recent literature has emphasized the influence that question format can have on the elicitation process, with specific reference to *interval* elicitation.

2.1. Overconfidence in Interval Judgments

2.1.1. One-, Two-, and 3-Point Interval Elicitation

Soll and Klayman[2] found that the way in which the interval was requested—the elicitation

[4] Other means of measuring interval calibration are available, for example, Yaniv and Foster's[5] L score measure of the tradeoff between accuracy and informativeness. However the measurements of hit rates predominate in the literature, making direct comparisons between different question formats far easier. In addition, Yaniv and Foster's trade-off formula can be biased toward narrow intervals far from the truth,[6] and has to be carefully adjusted on a question-by-question basis, making L score comparisons only meaningful between participants for a given question.

format—can substantially influence overconfidence. When intervals were elicited in the range format overconfidence was at its worst, whereas when intervals were elicited using the "3-point format" overconfidence was less (see Table I for an example of formats and corresponding measures of overconfidence). This approach is further supported by knowledge sampling theory,[13] which suggests that the more times we sample our knowledge base to create an estimate, the less overconfident that estimate is likely to be. The 4-step[5] procedure used in our studies combines this 3-point question format with participant assignment of confidence levels[3] discussed next.

2.1.2. Assigned Versus Anticipated Confidence

Teigen and Jorgensen[3] found that when participants were allowed to assign their own confidence levels, overconfidence was reduced. Across a series of five experiments, participants were asked to either provide an interval at a given level of confidence (where 90% intervals yielded 52% hit rates), or assign a confidence level to a preexisting interval (where 45% intervals yielded 44% hit rates). They found that when a confidence level was assigned to a preexisting interval overconfidence was substantially reduced. Judges appear to be better at assigning a reasonably accurate level of confidence to a given interval than to nominating an interval corresponding to a target assigned level of confidence.

2.1.3. Inclusion Versus Exclusion Intervals

Yaniv and Schul[14] compared the effects of *inclusion* and *exclusion* elicitation strategies using general knowledge questions with students ($n = 91$). Each question had 20 possible responses and participants were instructed to either highlight all the likely (inclusion), or unlikely answers (exclusion). They found that the inclusion instruction produced much smaller sets of responses (18–21% of the full set) than the exclusion instruction (43–50%). The exclusion instruction appeared to reduce overconfidence, improving the hit-and-miss calibration of the participants from 31–40% to 58–69%.

[5] We use the term "step" instead of "point" because unlike the 3-point question format the 4-step procedure elicits three point estimates, as well as a confidence level. The confidence level influences the overall width of the estimate and as such represents more than just another point in the interval distribution.

2.1.4. Expert Versus Novice Intervals

McKenzie *et al.*[15] compared intervals produced by novices and experts for the same set of questions. Two sets of participants ($n = 92$ UCSD students and $n = 43$ IT professionals) were asked two sets of questions (20 questions regarding UCSD and 20 questions regarding IT). In this way each participant answered questions in the role of both expert and novice. They found that experts and novices had similar levels of overconfidence (46% vs. 49%); however, the two groups were overconfident in different ways. Experts produced narrow intervals with the midpoint closer to the truth, whereas novices produced wide intervals with the midpoint further from truth.

2.2. How Does the 4-Step Procedure Reduce Overconfidence?

Although there is no single view as to theoretical underpinnings of reducing overconfidence in interval elicitation, Klayman *et al.*[13] suggest that there may be at least two independent mechanisms in play. In Reference 13, Soll and Klayman propose their knowledge sampling theory, that, is by asking several questions, a judge can assess several different hypotheses in constructing an interval estimate. For example, a judge asked to estimate Barack Obama's age when he became president (47 years old) may first take into account Obama's youthful appearance to consider his lowest possible age. To estimate Barack's highest plausible age, the judge may take into account the typical length of political career for a presidential candidate. In deciding his or her best estimate, a judge may weigh the evidence considered in determining the high and low estimates, or alternatively the judge may take into account new data (e.g., the age of Barack's children) in calculating his or her best estimate.

Having produced an interval, the 4-step procedure then asks the judge to review the interval and estimate the judge's level confidence that it captures the true value. In Reference 13, Juslin and Winman suggest that judges are much better at interval evaluation than at interval production, as do Teigan and Jorgensen.[3] The fourth step of the procedure therefore takes advantage of this observation, and offers the judge the opportunity to switch from interval production to interval evaluation when assigning a confidence level.

In combining these two independent views on the cognitive interpretation of interval estimates, we hope to offer a plausible theoretical framework as to how the 4-step procedure reduces overconfidence. We acknowledge that the ease with which a judge can switch between interval production and assessment tasks is untested, and concede this as a potential weakness with this explanation. In this article, however, our primary focus is on the empirical study of interval elicitation using expert groups, and details of studies we have conducted now follow.

3. EMPIRICAL STUDIES

The domains of infectious disease and marine ecology were chosen because of the availability of experts, and because the underlying systems are complex and data are often deficient, making expert judgments challenging. For example, the variety of pathways by which infectious diseases are spread and detected make forecasting difficult. Some diseases are blood borne (e.g., hepatitis B), or spread through contaminated food (e.g., salmonellosis) or water supplies (e.g., giardiasis). Others may be carried by an external vector, such as a mosquito (e.g., Ross River fever), or spread through airborne transmission of droplets from sneezes and coughs (e.g., influenza). With the addition of seasonal variations, rates for different diseases for a given population are not simple predictions. Expert judgments in marine ecology are similarly challenging. Population responses and physiological pathways of effects are typically poorly understood. Ecological responses to pollution are mitigated by seasonal variations in tides, proximity to international shipping, and interactions with other species.

In each of our studies we employed a 4-step elicitation procedure as summarized in Fig. 1. This combines two approaches that have been previously been successful in reducing overconfidence. That is, the 3-point question format (as used by Soll and Klayman[2]) and anticipated rather than assigned confidence levels (as used by Teigen and Jorgensen[3]).

In Study 1, the 4-step procedure was developed and tested on infectious disease experts, with the goal of minimizing overconfidence while exploring the relationship between overconfidence, cognitive style,[12] and forecast period. The elicitation procedure produced minimal overconfidence. So Studies 2 and 3 were conducted with different expert groups, using a randomized control group design, so that a direct comparison could be made between the 4-step procedure and the 3-point question format.[2] Study 3 had the additional goal of replicating the 4-step procedure in a different expert domain, namely, marine invertebrate populations.

It is important to distinguish between 95% confidence intervals[6] (CIs) that we calculate and report for our data, and the elicited intervals, which have an associated level of confidence, for example, 80%. The 95% data CIs represent a range of plausible values for the true mean proportion. In other words, about 95 of 100 CIs generated from the same population and process would capture the true population value. The elicited interval estimates represent the lower and upper limits of an interval, as estimated by an expert that in her opinion have an 80% chance of including the true value. For intervals where the expert assigned his or her own level of confidence, the corresponding 80% intervals, which we refer to as "derived" intervals, were calculated for the purpose of comparing estimates.

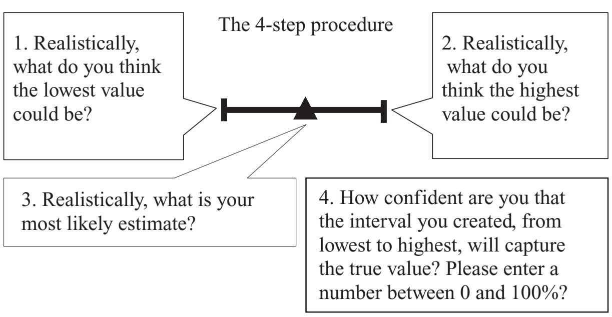


Fig. 1. The 4-step interval elicitation procedure.

3.1. Infectious Disease Experts—Study 1

3.1.1. Method

3.1.1.1. Participants. The participants ($n = 21$) were Australian infectious disease experts with a

[6] Throughout this article, empirical results are reported as proportions with 95% CIs. The use of CIs is less common than the use of null hypothesis testing and p values so a brief justification follows. Cumming and Finch[16] list four advantages of using CIs: they give point interval estimates in easily comprehensible units; p values can be easily estimated using CIs, with $\frac{1}{4}$ overlap of 95% CI width being approximately equivalent to $p<.05$ (for independent groups); CIs help combine evidence from multiple studies, thus facilitating meta-analytic thinking; and CIs provide information about the precision of effect estimates, and this can be regarded as taking the role of statistical power (wide CIs indicate less precision, narrow CIs relatively more). The following convention has been used for the in-text presentation of CIs in this article: mean value [lower limit of 95% CI, upper limit of 95% CI].

Giardiasis

Disease rates for the past five years

Year	2002	2003	2004	2005	2006
Rate of disease (100,000/annum)	14.6	15.8	15.6	18.5	23.8

1. Compared to the last five years, what do you think will happen to the rate of disease for the 12 months to September 30th 2007?

   rate will decrease rate will remain stable rate will increase

2. Realistically, what do you think the lowest rate of this disease could be for the 12 months to September 30th 2007? ______
3. Realistically, what do you think the highest rate of this disease could be for the 12 months to September 30th 2007? ______
4. Realistically, what is your best guess (i.e. most likely estimate) of the rate of disease for the 12 months to September 30th 2007? ______
5. How confident are you that your interval, from lowest to highest, could capture the reported rate for this disease? Please enter a number between 0 and 100%. ______

Fig. 2. An example question from Study 1, starting with the direction of change question, and followed by the 4-step interval elicitation procedure. Experts were canvassed during the period June to July 2007.

minimum of 5 years' relevant work experience and/or postgraduate qualification in a relevant field and currently working in a related government or university department, or private sector organization.

Participants were recruited via publicly available e-mail addresses obtained from relevant publications and websites. Of the 201 experts initially contacted, 13 did not meet the participant criteria. Of the 188 remaining, 48 responded—a response rate of 26%. However, 27 of those returned incomplete questionnaires, which were discarded because their forecast performance could not be adequately assessed. We acknowledge this is high attrition, probably caused by the survey being reasonably long (over 30 minutes) and repetitive (same questions for 10 diseases, two time periods), and extrinsic motivation being low (the chance to win a US$ 50 book voucher). The final response rate was 11%, and the demographics outlined below pertain to this 11% of respondents.

The average age of respondents was 44.6 years, with an average of 15.3 years of relevant work experience. The majority (62%) of participants were male. Almost all (95%) had postgraduate qualifications, with 33% holding Ph.D.s. One-third (33%) were employed in universities, 62% in government departments, and 5% in the private sector. The majority (81%) had published work in relevant peer-reviewed journals, and 71% were members of relevant professional organizations or societies.

3.1.1.2. Materials. A purpose built rates-of-disease questionnaire was used to elicit interval estimates. For each of the 10 diseases, five questions were asked regarding 2007 forecasts and five for 2012 forecasts. Fig. 2 presents the question format for the 2007 forecast period. Following this, demographic items were collected: age, gender, education level (Doctorate, Masters, Bachelor), relevant work experience (in years), institutional affiliation (university, government, private sector), publications in related journals (yes or no), membership of a related professional body (yes or no), and areas of expertise (a check-box for each disease). We also administered a styles-of-reasoning questionnaire (adapted from Tetlock[(12)]) to assess cognitive style; however, this is not the focus of this article.

3.1.1.3. Procedure. Participants were encouraged to complete all the questions in one sitting, as quickly as possible (~30 minutes). They were asked to predict the rates of 10 reportable diseases for both 2007 and 2012. The data presented here are for 2007 estimates; the analysis associated with the 2012 estimates is not reported on.[7] The 10 diseases were: cryptosporidiosis, giardiasis, listeriosis, salmonellosis, shigellosis,

[7] The 2012 results are not reported on because at the time of writing it was impossible to calibrate these results relative to true values.

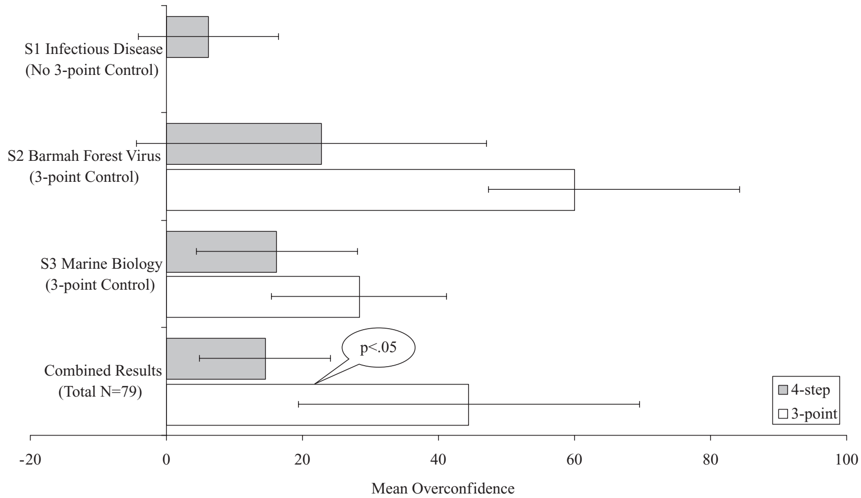


Fig. 3. The average overconfidence for all three studies, with comparisons between the 3-point question format and 4-step procedure for Studies 2 and 3. For intervals where the expert assigned his or her own level of confidence, the corresponding 80% intervals, which we refer to as "derived" intervals, were calculated for the purpose of comparing estimates. No conversion was necessary to compare participants' 3-point intervals as they were all elicited at an 80% level of confidence. Error bars represent 95% CIs.

Barmah Forest virus infection, dengue fever, flavivirus infection, malaria, and Ross River fever.

Rates of each disease for Victoria Australia for the last five years were provided to the participants. All infectious disease data (including the October 1, 2007 data used to calibrate expert predictions) were sourced from the Victorian Government Health Information website (http://www.health.vic.gov.au/ideas/surveillance/index.htm). For each disease, for each forecast period, participants were asked to predict the direction of change and then estimate: (1) the lower limit of the rate; (2) the upper limit of the rate; (3) their best guess of the rate; and (4) their confidence in the interval they had created (from lower to upper). The direction of change question was modeled on Tetlock's[12] survey design, and was included in Study 1 to ease participants into the elicitation process. In Study 3, however, this question was omitted because qualitative feedback from Study 1 suggested that we needed to minimize the number of questions per questionnaire and the direction of change question is redundant in that it elicits no additional information.

Estimates, in the form of a proportion out of 100,000, were subjected to an arcsine transformation[17] to reduce the asymmetry of score distributions. This is especially useful when, as here, values are very close to zero. Rates of disease are by convention reported as cases per 100,000 of population, and the rates of reportable diseases are typically low, in our case between 0 and 35 cases per 100,000. The arcsine transformation moves points at the extremes of the scale further from the mean value, and helps to ensure that subsequent statistical analysis where data with a normal distribution are assumed is not invalidated. In our case, each expert's original intervals were transformed to "derived" 80% confidence intervals to enable comparison between expert predictions. For example, participant no. 9 provided a 60% interval for the rate of giardiasis of 14–50 cases per 100,000. This was used to derive to a transformed 80% interval of 8.2–63.0, assuming a normal distribution of arcsine values. Forecast performance is reported here as the average overconfidence observed in experts' 80% derived intervals. For example, using the 80% derived intervals participant no. 9 produced 10 intervals that included the true value (i.e., "Hits"). Participant no. 9 is therefore considered to be 20% underconfident given her hit rate of 10/10.

3.1.2. Results

Average hit-and-miss calibration for the 2007 forecasts (using the derived 80% confidence level intervals) was 73.8%, 95% CI [63.5, 84.1], confirming that the 4-step procedure introduced minimal overconfidence, that is, $80\% - 73.8\% = 6.2\%$ overconfidence (see Fig. 3). Of the 21 experts, four provided derived intervals that included the true rate of disease for all 10 diseases (hit rate = 10/10). A further three experts achieved perfect calibration (hit rate = 8/10). Only one expert had zero hits.

Given that our response rate was low (11%), there is a potential problem with sample bias, namely, whether those experts who chose to respond are different in relevant and important ways from those who chose not to. For example, perhaps those who responded were more confident than those who did not. It seems plausible that those experts who think they know the answer to a question or who believe they can supply a reliable estimate are more likely to respond. This may mean that our sample self-selected as a more confident sample, more sensitive to question format. If the sample was biased to being especially overconfident this may not detract from the demonstrated effectiveness of our elicitation method. One possible interpretation is that the sample bias in fact serves to highlight the effectiveness of our method in reducing overconfidence. However, we do appreciate that other less complementary explanations are possible, and addressing this sample bias was a strong motivation for conducting the randomized control Studies 2 and 3. In these, a direct comparison was made between the performance of the 4-step and the 3-point question formats using samples drawn from workshop participants with varying levels of expertise.

3.2. Public Health Professionals—Study 2

Study 2 was designed to replicate Study 1 with the addition of a control group to compare the performance of the 4-step procedure with the 3-point question format.

3.2.1. Method

3.2.1.1. Participants. The participants ($n = 30$) were a group of public health professionals familiar with the biology and ecology of the Barmah Forest virus, but not with the reported rates of this disease over recent years. Participants were taking part in a risk analysis workshop in Canberra, Australia.

Participants had an average 8.1 years of relevant work experience. The majority (67%) of participants were female. Half the participants (50%) had postgraduate qualifications, with 8% holding Ph.D.s. All participants were employed in government departments. Participants came from a mix of professional backgrounds, including epidemiology (8%), health professionals (8%), infectious disease (25%), environmental toxicology (13%), pharmacology (8%), program development (17%), and public health (21%). Jelly beans were offered as a participation incentive.

3.2.1.2. Materials. The participants completed a paper questionnaire that included demographic questions, two paragraphs of background information on the Barmah Forest virus, the average rate of disease for 1997–2001, and a single forecast question regarding the average rate of disease over the period 2002–2006. Two versions of the questionnaire were randomly distributed. The two versions were the same, except that one used a 3-point question format and the other the 4-step procedure. Each participant saw only a single questionnaire.

3.2.1.3. Procedure. Participants were first advised that the average yearly notification rate (per 100,000 of population) for Barmah Forest virus infection in Australia for the years 1997–2001 was 3.9. They were then asked to produce a single interval estimate of the change in the average yearly notification rate of Barmah Forest virus infection in Australia between 1997–2001 and 2002–2006. One group ($n = 14$) used the 4-step procedure (i.e., the same format as Study 1), whereas the other ($n = 10$) used the 3-point question format[2] with an assigned 80% confidence level (see Fig. 4). The 3-point intervals were then compared with the 80% derived confidence intervals calculated from responses to the 4-step question format.

3.2.2. Results

The average hit-and-miss rate of the derived 80% intervals produced by the 4-step procedure was 57.1% [29.8, 81.4], compared with just 20.0% [5.7, 51.0] for the 3-point question format. In other words, the average overconfidence for the 4-step procedure was 22.9%, whereas participants using the 3-point question format were on average 60% overconfident (see Fig. 3).

3.3. Natural Resource Management—Study 3

Study 3 was designed to replicate Study 2 in a different expert domain. The opportunity was also taken to further refine the 4-step procedure by omitting the "direction of change" question included in Studies 1 and 2.

3.3.1. Method

3.3.1.1. Participants. The participants ($n = 34$) in this study were attendees at the Bayesian Network Modellers Meeting, Brisbane, November 19–20, 2007. This study formed only a small part of the two-day agenda. All workshop participants agreed to take part in the study, with 15 participants

The average yearly notification rate (per 100,000 of population) for Barmah Forest virus infection in Australia for the years 1997-2001 was 3.9.

Knowing this, we would like you to complete the questions in the table below.

In your opinion, the average yearly notification rate for Barmah Forest virus infection in Australia from 1997-2001 to 2002-2006 has:

Increased	decreased	remain stable
○	○	○

i. I am 90% confident the average yearly notification rate (per 100,000) for Barmah Forest virus for the period 2002-2006 is greater than	_____ per anum
ii. I am 90% confident the average yearly notification rate (per 100,000) for Barmah Forest virus for the period 2002-2006 is less than	_____ per anum
iii. Realistically, what is your best guess (i.e. most likely estimate) of the rate of disease for the 12 months to September 30th 2007?	_____ per anum

Fig. 4. The 3-point question format as used by the control group in Study 2.

completing the 3-point version of the questionnaire and 19 completing the 4-step questionnaire.

The average length of relevant work experience was 8.6 years. The majority (56%) of participants were male. Just over half the participants (53%) had postgraduate qualifications, with 47% holding Ph.D.s. All participants were affiliated with either universities (50%) or government (47%); one participant did not disclose his or her affiliation. All participants were in natural resource management roles but with a mix of professional backgrounds including ecotoxicology (6%); ecology (68%); and mathematical modeling (6%), 9% had more of a business orientation, and 12% did not disclose their professional background. A gift to the value of US$ 50 was offered to the individual giving the best responses. Of the three study groups, this group is considered the least expert in relation to the forecast questions. Although all participants had some insight into the impact of pollution on the population of invertebrates, only one considered himself or herself expert in this matter. No participants were previously familiar with the article used in the questionnaire.

3.3.1.2. Materials. Each participant received a handout containing case study information and a questionnaire containing either the 3-point or 4-step elicitation questions. The handouts contained excerpts from the introduction and methods section of a recently published journal article "Pollution Reduces Native Diversity and Increases Invader Dominance in Marine Hard-Substrate Communities."(18) The study investigated effects of different levels of pollution on marine sessile invertebrate species at four harbor sites in New South Wales, Australia.

The participants completed a paper questionnaire that included instructions, demographic questions, and eight forecast questions. An example of the 4-step question format (with the direction of change question omitted) is provided in Fig. 5. Questions 1a, 1b, 2a, and 2b provided participants with the mean number of species under different levels of pollution at two of the sites, and asked them to predict mean numbers of native and nonnative species for the two remaining experimental sites. Similarly, Questions 3a, 3b, 4a, and 4b provided participants with levels of percentage cover under the different levels of pollution at two sites, and asked them to predict percentage cover of native species for two levels of pollution at the remaining two sites.

3.3.1.3. Procedure. Participants were asked to produce eight intervals, predicting the population of marine invertebrates in polluted areas. The two versions of the questionnaire (3-point and 4-step) were randomly distributed between the participants. Questionnaires took roughly 30–45 minutes to complete, following which participants were provided with group feedback.

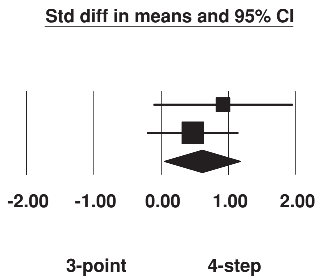
Unlike Studies 1 and 2, Questions 1a, 1b, 2a, and 2b in Study 3 asked for mean values rather than

1a.) The mean number of native species/plate on heavily polluted plates at Woolloomooloo Bay:	1b.) The mean number of non-indigenous species/plate on heavily polluted plates at Woolloomooloo Bay:
i. Realistically, the mean could be as low as ______ *species/plate*	i. Realistically, the mean could be as low as ______ *species/plate*
ii. Realistically, the mean could be as high as ______ *species/plate*	ii. Realistically, the mean could be as high as ______ *species/plate*
iii. Best guess of the mean is ______ *species/plate*	iii. Best guess of the mean is ______ *species/plate*
iv. For the interval I've created above from lower to upper), I think the chance that the mean observed in the study will fall in this interval is: (enter a number between 0 and 100) : ______ %	iv. For the interval I've created above from lower to upper), I think the chance that the mean observed in the study will fall in this interval is: (enter a number between 0 and 100) : ______ %

Fig. 5. The 4-step elicitation procedure as used in Study 3, without the direction of change question. Species/plate refers to the diversity of marine invertebrate species found on each 11 cm × 11 cm metal plate after being submerged in the ocean for seven months.

3-point Versus 4-step Overconfidence

Study name	Std diff in means	Standard error	Lower limit	Upper limit	p-Value
Study 2	0.920	0.529	-0.117	1.957	0.082
Study 3	0.470	0.346	-0.209	1.149	0.175
	0.605	0.290	0.037	1.173	0.037



Reducing Overconfidence in the Interval Judgments of Experts

Fig. 6. Meta-analysis reviewing the difference in effectiveness of the two-question formats combining the results of Studies 2 and 3. The combined results are highlighted in the dashed box and are expressed as a standardized difference in means—Cohen's *d*. These results suggest that the hit rate is improved by the 4-step procedure when compared to the 3-point question format.

proportions or rates, and so no arcsine transformation was used. The remaining questions in Study 3 (i.e., 3a, 3b, 4a, and 4b) used proportions (% cover) and so an arcsine transformation was applied. Derived 80% confidence intervals were calculated for the intervals elicited by the 4-step procedure for each question. The 80% derived intervals were then compared with the 80% confidence intervals elicited by the 3-point question format.

3.3.2. Results

The average hit-and-miss rate of the derived 80% intervals produced by the 4-step procedure was 63.8% [52.0, 75.6], compared to 51.7% [38.8, 64.5] for the 3-point question format. The average overconfidence using the 4-step procedure was 16.2%, whereas the average overconfidence for the 3-point question format was 28.3% (see Fig. 3).

3.4. Overall Results—Meta-Analysis

As a result of working with experts (as opposed to students), it proved challenging to recruit enough participants into any one study to establish high precision estimates of overconfidence. By combining the results of the three studies using meta-analytic techniques,[19] we can more precisely estimate the degree of overconfidence produced by the 4-step procedure. Mean overconfidence and the relevant standard error for the 4-step intervals for Studies 1, 2, and 3 were entered into comprehensive meta-analysis (CMA).[20] The weighted average overconfidence of the three studies, under a random effects model, was 11.9% [3.5, 20.3].

Fig. 6 shows a second meta-analysis (also using the random effects model), this time for the difference in hit rate between the 3-point and 4-step groups in Studies 2 and 3. The effect size (advantage of the 4-step condition) is displayed as Cohen's *d*, here indicating 0.6 of a standard deviation improvement in hit rate with the 4-step procedure as compared to the 3-point question format.

Additional meta-analysis was also conducted to assess if any of the demographic data were associated with improvements in hit rate when using the 4-step procedure across all three studies. A substantial effective size (Cohen's $d = 0.7$ [0.1, 1.3]) was found indicating that female participants were less overconfident when compared to male participants. This is consistent with findings elsewhere in the literature.[2,21] However, this finding is reported with some caution as a gender imbalance in Study 2 (3 males and 11 females) skews the results, and when Study 2 is removed from the analysis the effect size is diminished ($d = 0.5$ [−0.1, 1.2]).

4. DISCUSSION

The practical implication of this research for risk analysis is that simple changes to question format may substantially reduce overconfidence in the intervals elicited from experts. The combined results across the three studies suggest that the 4-step procedure produced minimal overconfidence of an average 11.9% [3.5, 20.3], and that a "hit" is more likely with 4-step intervals than with 3-point intervals. The advantage of 4-step intervals was present to varying degrees in each of the three expert groups (infectious disease, public health, natural resource management) and in two subject domains (infectious disease and marine ecology), as shown in Fig. 3.

In our studies to date, we have successfully implemented the 4-step procedure using both online and paper format questionnaires. In Studies 2 and 3 where we compared question formats, we encountered no additional resistance from experts responding to the 4-step procedure than to the 3-point format. We have received feedback that due to the repetitive nature of both question formats that fatigue does become an issue when responding to this type of questionnaire. We speculate that asking for 20 estimates in Study 1 (in addition to cognitive style and demographic questions) may have been too demanding, contributing to 27 out 48 responses being incomplete. However, in Study 3 when we asked for eight estimates (in addition to demographic questions), we received no incomplete responses. We therefore suggest as a rule-of-thumb for anyone planning to develop an elicitation questionnaire using the 4-step procedure that an optimal number of estimates to be elicited in any one questionnaire be between 8 and 12.

4.1. Limitations

We acknowledge that an extra analysis step is required to transform participant 4-step intervals—of different confidence levels—to derived 80% confidence intervals for comparative analysis. In addition, this transformation function should be appropriate for the type of data being elicited, for example, arcsine for proportions. However, this step can be easily automated, in our case using simple formulas in Excel, after which it becomes transparent to the data analysis process.

A major challenge in working with expert populations is that sample sizes may well be limited, resulting in low precision manifested as wide CIs. In this article, we have endeavored to address this by aggregating the results of three studies using meta-analytic techniques. Meta-analysis is an established method for aggregating the results of studies with low power to find small but meaningful effects.[19] To have observed a difference of comparable magnitude across three separate experiments in diverse expert domains seems to us very good evidence of a consistent and robust effect. One could argue that this is stronger evidence even than that which might be gathered by a larger single trial.

4.2. Future Directions

Experts sometimes assigned very low confidence to their intervals (e.g., 30%, 50%). When these very low confidence intervals are transformed back into the corresponding derived 80% or 90% intervals, they were consequently extremely wide. This improved the hit rate of the interval, at the expense of estimate precision.[5] We speculate that if experts saw the width of their derived intervals, they themselves would often not accept the lack of precision, that is, excessive interval width.

Initial refinement in the question wording of the 4-step has included the limiting the assigned level of confidence to between 50 and 100%. Previously we allowed values between 0% and 100%. Assigning a confidence level of less than 50% implies that the true value is more likely to be outside of the specified interval than within it. For this reason, this simple modification to the 4-step procedure has now been adopted.

In addition, we plan to investigate the effect of feeding back to experts their derived intervals and will measure the extent of adjustments they make to these intervals and/or to their confidence levels. Visual representations—figures, diagrams, charts—have been shown to be most effective, when used appropriately, in conveying probabilistic and statistical information.[22] Visual interactive displays may therefore help improve the process of interval elicitation by clarifying for experts the relationship between confidence and precision. We will investigate how visual interactive feedback changes interval precision and whether any desirable reduction in overconfidence produced by the 4-step elicitation method survives such adjustments.

Further investigation into the theoretical mechanism underpinning the 4-step procedure will also be facilitated by interactive questionnaires. With appropriate instrumentation we will be able to gain an enhanced understanding of how an expert constructs and adjusts an interval estimate. Will they choose to adjust the limits of their estimate, or will they increase or reduce their assigned confidence level? Perhaps individual preferences will be split and so by offering two adjustment mechanisms we can reduce overconfidence in a greater proportion of the population. Instrumented interactive questionnaires will be key to answering these questions, and developing further insight into the cognitive basis of overconfidence.

A second solution to the problem of wide-derived intervals is to explore alternative methods for adjusting expert intervals that do not entail the assumption of normality when transforming the expert-nominated confidence level to the derived 80% confidence level. Perhaps lognormal or other transforms may better reflect experts' underlying subjective distributions, and these will be explored in further research.

5. CONCLUSION

Overconfidence has an adverse impact on the process of risk analysis, and is influenced by many factors. Although cognitive heuristics and biases can prove challenging to correct,[23,24] refinements in elicitation procedure and question format can be easily implemented. In summary, using the 4-step elicitation procedure may help experts more accurately express the extent of their uncertainty about an elicited quantity.

ACKNOWLEDGMENTS

The authors thank everyone for their words of encouragement while writing this article. This work was funded by ACERA Project 0611 and NSF Award SES-0725025 to Decision Research.

REFERENCES

1. Onkal D, Muradoglu G. Effects of task format on the probabilistic forecasting of stock prices. International Journal of Forecasting, 1996; 12:9–24.
2. Soll JB, Klayman J. Overconfidence in interval estimates. Journal of Experimental Psychology Learning Memory and Cognition, 2004; 30(2):299–314.
3. Teigen KH, Jorgensen M. When 90% confidence intervals are only 50% certain: On credibility of credible intervals. Applied Cognitive Psychology, 2005; 19:455–475.
4. Morgan MG, Keith DW. Subjective judgments by climate experts. Environmental Science & Technology, 1995; 29(10):468–476.
5. Yaniv I, Foster DP. Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. Journal of Experimental Psychology, 1995; 124(4):424–432.
6. Speirs-Bridge A. The influence of elicitation format, cognitive style, and forecast period on the rates-of-disease predictions of infectious disease experts (Honours Thesis), 2007. Available at: http://members.optusnet.com.au/speirsbridge/psych/Speirs-BridgeThesisHonours2007.pdf, Accessed on October 26, 2009.
7. Oskamp S. Overconfidence in case-study judgments. Journal of Consulting Psychology, 1965; 29:261–265.
8. Tsai CI, Klayman J, Hastie R. Effects of amount of information on judgment accuracy and confidence. Organizational Behavior and Human Decision Processes, 2008; 107(2):97–105.
9. Lichtenstein S, Fischhoff B, Phillips L. Calibration of probabilities: The state of the art 1980. Pp. 306–334 in Kahneman D, Slovic P, Tversky A (eds). Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1982.
10. Lichtenstein S, Fischhoff B. Do those who know more also know more about how much they know? Organizational Behaviour and Human Decision Processes, 1977; 20:159–183.
11. Dawes RM. House of Cards: Psychology and Psychotherapy Built on Myth. New York: Free Press, 1994.
12. Tetlock PE. Expert Political Judgment. Princeton, NJ: Princeton University Press, 2005.
13. Klayman J, Soll JB, Juslin P. Winman A. Subjective confidence and sampling of knowledge. Pp. 153–182 in Fiedler K, Juslin P (eds). Information Sampling and Adaptive. New York: Cambridge University Press, 2006.
14. Yaniv I, Schul Y. Elimination and inclusion procedures in judgment. Journal of Behavioral Decision Making, 1997; 10(3):211–220.
15. McKenzie CRM, Liersch MJ, Yaniv, I. Overconfidence in interval estimates: What does expertise buy you? 2008. Organizational Behavior and Human Decision Processes, 2008; 107(2):179–191
16. Cumming G, Finch S. Inference by eye: Confidence intervals, and how to read pictures of data. American Psychologist, 2005; 60:170–180.
17. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences, 2nd ed. Belmont, CA: Brooks Cole, 1982.

18. Piola RF, Johnston EL. Pollution reduces native diversity and increases invader dominance in marine hard-substrate communities. Diversity and Distributions, Journal Compilation, 2008; 14(2):329–342.
19. Borenstein M, Hedges L, Higgins J, Rothstein H. Introduction to Meta-Analysis. Chichester, UK: John Wiley & Sons Ltd., 2009.
20. Borenstein M, Hedges L, Higgins, J, Rothstein H. Comprehensive Meta-Analysis Version 2 [Computer Program]. Englewood, NJ: Biostat, 2005.
21. Barber BM, Odean T. Boys will be boys: Gender, overconfidence, and common stock investment. Quarterly Journal of Economics, 2001; 116(1):261–292.
22. Tufte ER. The Visual Display of Quantitative Information. Cheshire, CT: Graphics Press, 1983.
23. Tversky A, Kahneman D (eds). Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1982.
24. Plous S. The Psychology of Judgment and Decision Making. New York: McGraw-Hill Inc., 1993.