Report Cover Page

ACERA Project
0611
Title
Eliciting reliable expert judgements for ecological models
Author(s) / Address (es)
Andrew Speirs-Bridge[1], Fiona Fidler[1,2], Marissa McBride[2], Louisa Flander[3], Geoff Cumming[1], Mark Burgman[2] 1. School of Psychological Science, La Trobe University, Australia. 2. Australian Centre for Excellence in Risk Analysis (ACERA), University of Melbourne, Australia. 3. School of Population Health, University of Melbourne, Australia.
Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)
Manuscript
Summary
This report outlines a new method for eliciting judgements from experts. Based on recent developments in experimental psychology, it applies a standardised method to three different situations. The results indicate that the method substantially reduces expert over-confidence, leading to better calibrated judgements.

ACERA Use only	Received By:	Date:
	ACERA / AMSI SAC Approval:	Date:
	DAFF Endorsement: ( ) Yes ( ) No	Date:

Eliciting reliable expert judgements for ecological models; ACERA Project No. 0611

Mark Burgman
University of Melbourne

Reducing Overconfidence in the Interval Judgments of Experts

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

1. School of Psychological Science, La Trobe University, Australia.
2. Australian Centre for Excellence in Risk Analysis (ACERA), University of Melbourne, Australia.
3. School of Population Health, University of Melbourne, Australia.

Manuscript (final report)

July 2008

Acknowledgements

This report is a product of the Australian Centre of Excellence for Risk Analysis (ACERA). In preparing this report, the authors acknowledge the financial and other support provided by the Department of Agriculture, Fisheries and Forestry (DAFF), the University of Melbourne, Australian Mathematical Sciences Institute (AMSI) and Australian Research Centre for Urban Ecology (ARCUE).

Disclaimer

This report has been prepared by consultants for the Australian Centre of Excellence for Risk Analysis (ACERA) and the views expressed do not necessarily reflect those of ACERA. ACERA cannot guarantee the accuracy of the report, and does not accept liability for any loss or damage incurred as a result of relying on its accuracy.

Table of contents

Acknowledgements 3
Disclaimer 4
Table of contents 5
List of Tables 6
List of Figures 7
Executive Summary 8
Introduction 8
Overconfidence: A brief overview 9
Empirical Studies 11
Study 1: Infectious Disease Experts 13
Study 2: Public Health Professionals 16
Study 3: Natural Resource Management 18
Overall Results: Meta-analysis 20
Discussion 21
References 22

List of Tables

Table 1..... 9

List of Figures

Figure 1..... 11
Figure 2 14
Figure 3 15
Figure 4 16
Figure 5..... 19
Figure 6 20

Executive Summary

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, creating an interval of some assigned confidence level (e.g., 80%). In our 4-step interval elicitation procedure experts were also asked for a lower limit, upper limit and best guess; the fourth step was rating their expected confidence in the interval produced. In our three studies, experts made interval predictions of rates of infectious diseases (Studies 1, *n*=21, and 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 compared with previous studies. Studies 2 and 3 suggest the 4-step procedure is more likely to produce a hit than the 3-point procedure (Cohen's *d*=0.61, 95%CI [0.04, 1.18]).

Introduction

Expert elicitation forms a key step in risk analysis. Furthermore the critical role of question format in the elicitation of judgments is well supported by the research literature [1, 2 & 3]. It is possible to achieve substantial reduction in overconfidence by refining the question format—for example: Soll and Klayman [2] reduced overconfidence by 27% using a 3-step interval elicitation procedure; and Teigen and Jorgensen [3] reduced it by 52% using 'expected' confidence levels rather than 'assigned' confidence levels. To date, however, these reductions have been observed only in student samples for general knowledge questions. Our studies investigate: (a) whether these impressive reductions in overconfidence can be replicated in expert samples and (b) if further reductions in overconfidence can be achieved by combining the methods of Soll and Klayman with those of Teigen and Jorgensen.

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 (e.g. [4, 5 & 6]). 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. By 'hit-and-miss calibration' we mean that if an expert provided ten 80% confidence intervals, 8 of those 10 should contain the truth for the expert to be perfectly calibrated. If instead the expert's intervals only contained the truth 6 of 10 times then we consider them 20% *overconfident*.

Overconfidence: A brief overview

There is a vast 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, 10 & 11]; and can be influenced by the experts' *cognitive style* [12]. Whilst 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 emphasised the influence that question format can have on the elicitation process, with specific reference to *interval* elicitation.

Overconfidence in interval judgments

1, 2 and 3 point interval elicitation: Soll and Klayman [2] found that the way in which the interval was requested—the elicitation 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 procedure used in our studies combines this 3-point question format with participant assignment of confidence levels [3] discussed next.

***Table I**. Overconfidence, Expressed As Confidence Level Minus Hit Rate.*

Elicitation Format with Examples	Average Overconfidence
Following Soll and Klayman [2].	
1-point (range)	41%
I am 80% sure that this happened between _____ and _____.	
2-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 it's equally likely that this happened after or before _____.	
Following 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 (Free) Confidence Level	15%
The year Einstein was born is between _____ and _____. I am _____% confident that my interval contains the true answer.	

Assigned versus expected 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 predetermined interval (where 45% intervals yielded 44% hit rates). They found that when a confidence level was assigned to a pre-existing 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.

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%.

Expert versus novice intervals. McKenzie, Liersch and Yaniv [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% versus 49%), however each group was overconfident for different reasons. Novices produced wide intervals with the midpoint further from truth, whereas experts produced narrow intervals where midpoint was closer to the truth.

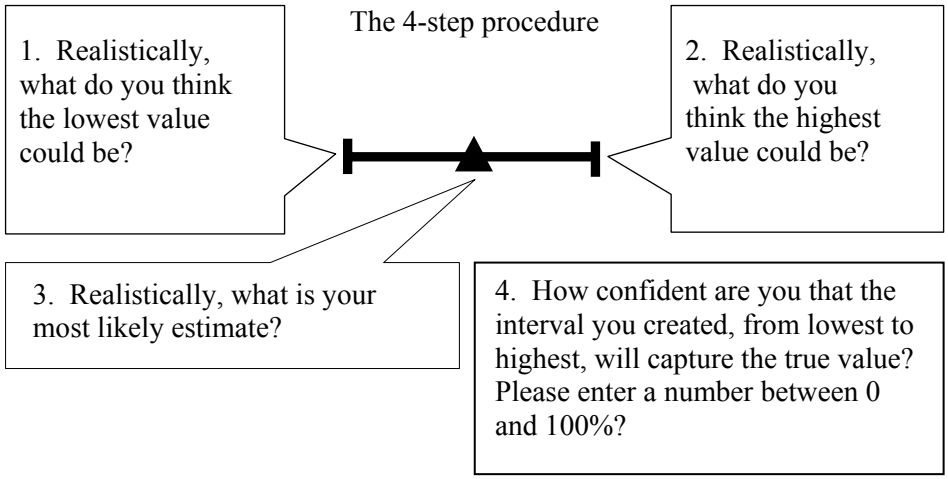
Empirical Studies

Rationale

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 summarised in Figure 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 expected rather than assigned confidence levels (as used by Teigen and Jorgensen [3]).

Figure 1*. The 4-step interval elicitation procedure.*



In Study 1 the 4-step procedure was developed and tested on infectious disease experts, with the goal of minimising overconfidence whilst exploring the relationship between overconfidence, styles-of-reasoning, and forecast period. The elicitation procedure produced minimal overconfidence so Studies 2 and 3 were conducted with different expert groups, using a randomised 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.

Throughout this article, empirical results are reported as proportions with 95% confidence intervals (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 ¼ 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 paper: mean value [lower limit of 95% CI, upper limit of 95% CI].

It is important to distinguish between the calculated 95% CIs and the elicited interval estimates. The 95% 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 lowest and highest possible values, as predicted by an expert, that in her opinion have an 80% chance of including the true value. For intervals where the expert assigned their own level of confidence, the corresponding 80% intervals, which we refer to as 'derived' intervals, were calculated for the purpose of comparing estimates with true values.

Study 1: Infectious Disease Experts

Method

Participants. The participants (*n*=21) were Australian infectious disease experts with a 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 organisation. Participants who did not meet these criteria were excluded from the study.

Participants were recruited via publicly available email 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 mins) and repetitive (same questions for 10 diseases, two time periods), and extrinsic motivation being low (the chance to win a $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 PhDs. 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 organisations or societies.

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. Figure 2 presents the question format for the 2007 forecast period. Following this, demographic items were collected: age (in years); 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]) however this is not the focus of this paper.

Figure 2. An example introductory question about direction of change followed by the steps of the 4-step interval elicitation method used in Studies 1 and 2. Experts were canvassed during the period June – July 2007.

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%. _____

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. The 10 diseases were: Cryptosporidiosis, Giardiasis, Listeriosis, Salmonellosis, Shigellosis, 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 1st 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 time period, participants were asked to predict the direction of expected change and then estimate 1). The lower limit of the expected rate; 2). The upper limit of the expected rate; 3). Their best guess of the expected rate; and 4). Their confidence in the interval they had created (from lower to upper).

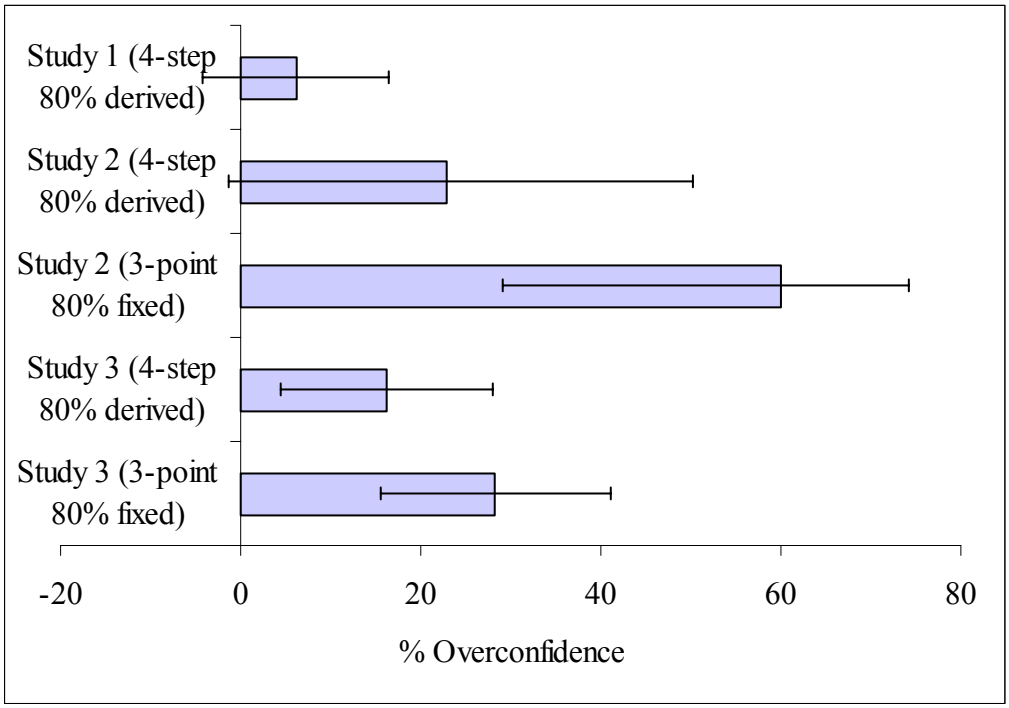
Estimates, in the form of a proportion out of 100,000, were subjected to an arcsine transformation [17] to reduce the asymmetry of distributions of values very close to zero. The experts' original 'free' intervals were then transformed to 'derived' 80% confidence intervals. For example, participant #2 originally provided a 30% 'free' interval for the rate of Giardiasis

disease from 20 to 27. This was used to derive to a transformed 80% interval of 13.2 to 36.5, 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 #2 produced 9 intervals that included the true value (i.e. 'Hits'), with only the interval for Cryptosporidiosis considered a 'Miss'. Participant #2 is therefore considered to be 10% underconfident given her hit rate of 9/10.

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, i.e., 80%-73.8%=6.2% overconfidence—see Figure 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.

Figure 3. The average overconfidence for all three studies, with comparisons between the 3-point and 4-step question formats for studies 2 and 3. Error bars represent 95% CIs.



Study 2: Public Health Professionals

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.

Method

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 PhDs. 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.

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.

Figure 4. An example introductory question about direction of change followed by an example of the 3-point question format, as used by the control group in Study 2.

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 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 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

Procedure. Experts were asked to produce one interval regarding their estimate of the change in the average yearly notification rate of Barmah Forest virus infection in Australia between 1997-2001 and 2002-2006. They were advised that The average yearly notification rate (per 100,000 population) for Barmah Forest virus infection in Australia for the years 1997-2001 was 3.9. 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 'fixed' 80% confidence level (see Figure 4). The uneven final group sizes (*n*=14 Vs *n*=10) may reflect a difference in willingness to participate in the two tasks. Almost all (14 of 15) participants provided responses in the 4-step question format, whereas only two thirds (10 of 15) were willing to provide responses for the 3-point question format. The 80% fixed 3-point intervals were then compared with the 80% derived confidence intervals calculated from responses to the 4-step question format.

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 fixed 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 Figure 3).

Study 3: Natural Resource Management

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 that was included in the previous two studies.

Method

Participants. The participants (*n*=34) in this study were attendees at the Bayesian Network Modellers Meeting (MBNM), Brisbane, 19-20 Nov 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 completing the 3-point version of the questionnaire and 19 completing the 4-step questionnaire.

The average years of relevant work experience was 8.6. The majority (56%) of participants were male. Just over half the participants (53%) had postgraduate qualifications, with 47% holding PhDs. All participants were affiliated with either universities (50%) or government (47%); one participant did not disclose their affiliation. All participants were in Natural Resource Management roles but with a mix of professional backgrounds including ecotoxicology (6%), ecology (68%); and mathematical modelling (6%), 9% had more of a business orientation, and 12% did not disclose their professional background. A gift to the value of $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. Whilst all participants had some insight into the impact of pollution on the population of invertebrates, only one considered themselves expert in this matter. No participants were previously familiar with the article used in the questionnaire.

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 harbour 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 (with the direction of change question omitted) is provided in Figure 5. Questions 1a, 1b, 2a and 2b provided participants with the mean number of species under different levels of pollution at 2 of the sites, and asked them to predict mean numbers of native and non-native species for the 2 remaining experimental sites. Similarly Questions 3a, 3b, 4a and 4b provided participants with levels of percentage cover under the different levels of pollution at 2 sites, and asked them to predict percentage cover of native species for 2 levels of pollution at the remaining 2 sites.

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

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: (type 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: (type a number between 0 and 100) : _____ %

Procedure. Participants were asked to produce eight intervals regarding 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 proportions or rates, and so no arcsine transformation was used. The remaining questions in Study 3 (i.e., 3a, 3b, 4a, and 4b) again 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 questions. The 80% derived intervals were then compared with the 80% confidence intervals elicited by the 3-point question format.

Results

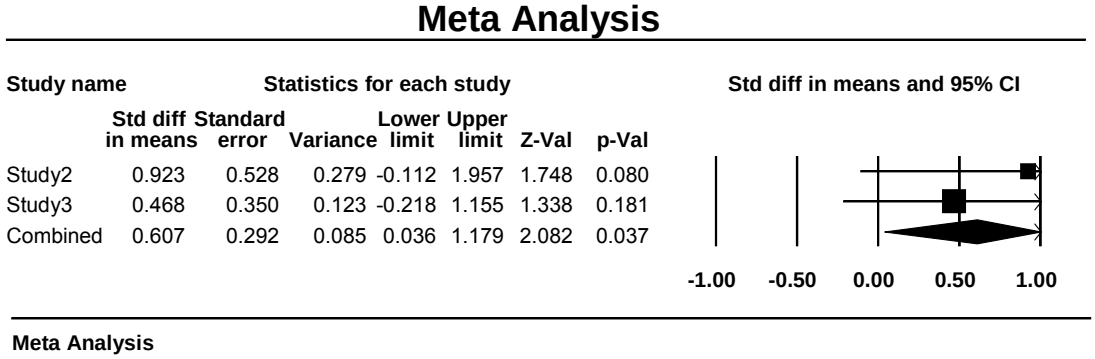
The average hit-and-miss rate of the derived 80% intervals produced by the 4-step procedure was 63.8% [52.0, 75.6], compared to 51.7% [38.8, 64.5] for the 3-point question format. In terms of overconfidence the average overconfidence using the 4-step procedure was 16.2%, whereas the average overconfidence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

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 sufficient precision to produce statistically significant results. By combining the results of the three studies using meta-analytic techniques 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 was entered into CMA (Comprehensive Meta-Analysis [19]). The weighted average overconfidence of the 3 studies, under a random effects model, was 11.9% [3.5, 20.3].

Figure 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.

Figure 6. Meta-analysis reviewing the difference in effectiveness of the two question formats combining the results of studies 2 and 3. The combined results (expressed in effect size—Cohen's *d*) suggest that hit rate is improved by the 4-step procedure as compared to the 3-point question format.



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).

Limitations

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 paper we have endeavoured to address this by aggregating the results of three studies using meta-analytic techniques.

Future directions

Experts sometimes assigned very low confidence levels to their intervals (e.g., 30%, 50%). When these very low confidence intervals are transformed back into the corresponding standard 80% or 90% intervals—which we referred to as 'derived' intervals—they were consequently extremely wide. This improved the hit rate of the interval, at the expense of estimate precision [20]. We speculate that if experts saw the width of their derived intervals, they themselves would often not accept the lack of precision, i.e. excessive interval width.

We propose to investigate next 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. In particular, we will investigate how interactive feedback changes interval precision and whether any desirable reduction in overconfidence produced by the 4-step elicitation method survives such adjustments.

Visual representations—figures, diagrams, charts—have been shown to be most effective, when used appropriately, in conveying probabilistic and statistical information [21]. Visual interactive displays may help improve the process of interval elicitation by clarifying for experts the relationship between confidence and precision.

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 expert's underlying subjective distributions, and these will be explored in further research.

Conclusion

Overconfidence has an adverse impact on the process of risk analysis, and is influenced by many factors. Whilst cognitive heuristics and biases can prove challenging to correct [22 & 23], 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 the uncertainty they have about some quantity.

References

Onkal, D., & Muradoglu, G. (1996). Effects of task format on the probabilistic forecasting stock prices. *International Journal of Forecasting, 12*, 9-24.

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

Teigen, K. H., & Jorgensen, M. (2005). When 90% confidence intervals are only 50% certain: On credibility of credible intervals. *Applied Cognitive Psychology, 19*, 455-475.

Morgan, M. G., & Keith, D. W. (1995). Subjective Judgments by Climate Experts. *Environmental Science & Technology, 29* (10), 468–476.

Van Der Fels-Klerx, H. J., Goossens, L. H. J., Saatkamp, H. W., Horst, S. H. S. (2002). Elicitation of Quantitative Data from a Heterogeneous Expert Panel: Formal Process and Application in Animal Health. *Risk Analysis, 22* (1), 67–81.

Cooke. R. M., & Goosens, L. H. J., 2004. Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research, 7* (6), 643–656.

Oskamp, S. (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology, 29*, 261-265.

Tsai, C. I., Klayman, J., & Hastie, R. (In Press). Effects of amount of information on judgment accuracy and confidence. *Organizational Behavior and Human Decision Processes*.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (306-334). New York: Cambridge University Press.

Lichtenstein, S. & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behaviour and Human Decision Processes, 20*, 159-183.

Dawes, R. M. (1994). *House of Cards: Psychology and Psychotherapy Built on Myth*. New York: Free Press.

Tetlock, P. E. (2005). *Expert Political Judgment*. New Jersey: Princeton University Press.

Klayman, J., Soll, J. B., Juslin, P., & Winman, A. (2006). Subjective confidence and sampling of knowledge. In K. Fiedler & P. Juslin (Eds), *Information Sampling and Adaptive Cognition* (153-182). New York: Cambridge University Press.

Yaniv, I., & Schul, Y. (1997). Elimination and Inclusion Procedures in Judgment. *Journal of Behavioral Decision Making, 10* (3), 211-220.

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in Interval Estimates: What does expertise buy you? Retrieved July 21, 2008, from http://management.ucsd.edu/faculty/directory/liersch/docs/overconfidence.pdf

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170-180.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks Cole.

Piola, R. F., & Johnston, E. L. (2008). Pollution reduces native diversity and increases invader dominance in marine hard-substrate communities. *Diversity and Distributions, Journal Compilation*, *14* (2), 329–342.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-Analysis Version 2 [Computer Program]. Englewood, NJ: Biostat.

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

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CN: GraphicsPress.

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

Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill Inc.