

Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research

Marissa F. McBride*, Fiona Fidler and Mark A. Burgman

Australian Centre of Excellence for Risk Analysis, School of Botany, University of Melbourne, Parkville, Vic. 3010, Australia

ABSTRACT

Aim Expert knowledge routinely informs ecological research and decision-making. Its reliability is often questioned, but is rarely subject to empirical testing and validation. We investigate the ability of experts to make quantitative predictions of variables for which the answers are known.

Location Global.

Methods Experts in four ecological subfields were asked to make predictions about the outcomes of scientific studies, in the form of unpublished (in press) journal articles, based on information in the article introduction and methods sections. Estimates from students were elicited for one case study for comparison. For each variable, participants assessed a lower and upper bound, best guess and level of confidence that the observed value will lie within their ascribed interval. Responses were assessed for (1) accuracy: the degree to which predictions corresponded with observed experimental results, (2) informativeness: precision of the uncertainty bounds, and (3) calibration: degree to which the uncertainty bounds contained the truth as often as specified.

Results Expert responses were found to be overconfident, specifying 80% confidence intervals that captured the truth only 49–65% of the time. In contrast, student 80% intervals captured the truth 76% of the time, displaying close to perfect calibration. Best estimates from experts were on average more accurate than those from students. The best students outperformed the worst experts. No consistent relationships were observed between performance and years of experience, publication record or self-assessment of expertise.

Main conclusions Experts possess valuable knowledge but may require training to communicate this knowledge accurately. Expert status is a poor guide to good performance. In the absence of training and information on past performance, simple averages of expert responses provide a robust counter to individual variation in performance.

Keywords

Calibration, expert elicitation, expert knowledge, overconfidence, subjective judgment, uncertainty.

*Correspondence: Marissa F. McBride, School of Botany, University of Melbourne, Parkville, Vic. 3010, Australia.
E-mail: mfc.mcbride@gmail.com

INTRODUCTION

Managers and decision-makers often rely on expert knowledge to inform policy and management decisions under uncertainty (Burgman, 2005; Sutherland, 2006). This may include facts or evidence recalled by the expert, inferences made by the expert to new or undocumented situations,

and/or the integration of disparate sources of information to address new problems (Kaplan, 1992). Experts may assist decisions by structuring a problem framework, building a conceptual model or selecting an analytical approach. They may also provide estimates of variables or event outcomes and their associated uncertainty. Particularly in cases when time or resources are limited, experts represent a critical,

alternative source of information for decision-makers (Kuhnert *et al.*, 2010).

Quantitative predictions from experts are of particular use for decision-making (Winkler, 1967). Expert estimates of facts and future events have been evaluated in many fields, including weather forecasting (Murphy & Winkler, 1984), accounting (Ashton, 1974), finance (Önköl *et al.*, 2003), clinical medicine (Christensen-Szalanski *et al.*, 1982), psychiatry (Oskamp, 1965) and engineering (Jorgensen *et al.*, 2004). Reliable good judgments have been reported for meteorologists and bridge players (Murphy & Winkler, 1977; Keren, 1987; Camerer & Johnson, 1991; Shanteau, 1992; Ericsson & Lehmann, 1996; Weiss *et al.*, 2006; McKenzie *et al.*, 2008); however, in most other disciplines, expert judgments are frequently shown to be inaccurate, biased and overconfident. Typically, metrics of expertise such as experience and level of training are only weakly related to performance (e.g. Camerer & Johnson, 1991; Burgman *et al.*, 2011). In some cases, experts perform the same as, or worse, than novices (Camerer & Johnson, 1991).

Expert skill is based on domain knowledge and repeated experience on relevant tasks with high-quality feedback (Ericsson & Kintsch, 1995; Ericsson, 2004). In domains where experts receive high levels of immediate, informative feedback, and there are incentives to improve, they learn to make accurate quantitative judgments (Hearst, 1988). Where these conditions are lacking, experts are likely to have difficulty making judgments (Fischhoff, 1990; Bolger, 1995). For example, ineffective environmental practices frequently persist simply because management agencies cannot be sure whether their successes or failures are a result of their actions (e.g. Hilborn & Ludwig, 1993; Sutherland, 2006; Roura-Pascual *et al.*, 2009).

While expert judgment is frequently used for ecological decision-making, few studies have directly investigated the validity of expert knowledge in ecological contexts (Iglesias & Kothmann, 1998; Johnson & Gillingham, 2004; Burgman *et al.*, 2005). Some expert judgments have been validated for species habitat models: in many of these cases, expert knowledge has improved model accuracy (e.g. Seoane *et al.*, 2005; Doswald *et al.*, 2007; Irvine *et al.*, 2009; Murray *et al.*, 2009), although this is not always the case (e.g. McCoy *et al.*, 1999; Pearce *et al.*, 2001; Lele & Allen, 2006). Bias (e.g. Crome *et al.*, 1996; Campbell, 2002; McCarthy *et al.*, 2004; Scholes & Biggs, 2006; Whitfield *et al.*, 2008) and errors (e.g. Stanley & Skagen, 2007; Kuhnert *et al.*, 2010) in judgments have been apparent in a number of ecological contexts. Very little work has been performed in ecology evaluating experts' assessments of uncertainty (but see Rothlisberger *et al.*, 2010). Perhaps the most consistent finding is the large, sometimes non-overlapping variation in estimates between experts (e.g. Morgan *et al.*, 2001; Yamada *et al.*, 2003; Czembor & Vesik, 2009).

The inconclusiveness and relative scarcity of evaluations of expert ecological knowledge represent a serious gap in applied ecology. The findings from other domains indicate that evaluation is necessary and that there is a role for a rigorous testing framework for expert knowledge. Furthermore, we

speculate that ecology does not offer the conditions for readily verifiable, timely, unambiguous feedback (Hilborn & Ludwig, 1993; Carpenter, 2002; Fazey *et al.*, 2005). In such cases, it is important to find out how experts actually perform, especially when their judgments are used to inform environmental policy and management decision-making.

In this study, we investigate the performance of ecological experts by asking them to make a range of quantitative predictions about the outcomes of scientific studies, in the form of unpublished (in press) journal articles. The questions test the ability of experts to extrapolate their existing knowledge to new settings within their field of expertise. We assess the degree to which expert predictions match observed experimental results and their ability to reliably assign uncertainty bounds to their estimates. We investigate relationships between experience, publication record, peer recognition and performance. The results provide insight into how ecological experts may be able to best inform environmental decisions.

METHODS

Defining 'good' judgment

Good quality judgments are those that 'exhibit a close correspondence with the observed outcome' (Murphy, 1993). When estimates from experts include uncertainty, we are interested in their:

1. Accuracy: the degree to which an expert's judgments correspond to the truth (external validity), and
2. Informativeness: the sharpness or level of precision with which a forecast predicts the true outcome (narrower confidence intervals indicate more informative responses).

High-quality forecasts will be those that are both accurate and informative [i.e. close to the truth and with justifiably confident (narrow) intervals]. Often there will be a trade-off between the two. For example, experts may specify overly wide bounds to guarantee including the truth.

Evaluation metrics

We focus on three aspects of response quality that have received considerable attention in the literature: (1) the calibration of confidence intervals, (2) interval informativeness, and (3) accuracy of point estimates.

Evaluation of interval calibration

In this study, an expert is well calibrated if, over the long run, for all predictions assigned a given probability the proportion that are true equals the probability assigned (Lichtenstein & Fischhoff, 1980). For example, if a well-calibrated expert provides estimates of lower and upper bounds (intervals) for a set of 10 quantities and states they are 80% confident that each interval will contain the true value, we expect the true value to fall within the bounds in 8 out of 10 cases. We refer to this

observed proportion of intervals that contain the truth as the ‘hit rate’. An interval with an expressed confidence level such as 80% is a subjective confidence interval (Savage, 1954). However, for brevity, we refer to ‘80% subjective confidence intervals’ as ‘80% intervals’ and use ‘confidence interval’ to refer to frequentist confidence intervals.

Evaluation of interval informativeness

The informativeness of a distribution measures how concentrated it is. Information must be assessed relative to some background distribution, usually the uniform or loguniform distribution. If expert i assesses probability distributions for N quantities, Cooke’s information score (Cooke, 1991) calculates the average relative information as

$$I_i = \frac{1}{N} \sum_{n=1}^N I(f_{i,n} | g_n), \quad (1)$$

where g_n is the background probability density for quantity n over the total range R_i^n and $f_{i,n}$ is the probability density function associated with expert i ’s response. The quantity $I(f | g)$ of f with respect to g is also known as the Kullback–Leibler distance. The total range for each variable is defined as

$$R_i^n = [q_{Ln}, q_{Un}], \quad (2)$$

such that the probability that the true value would fall outside the range $[q_{Ln}, q_{Un}]$ is expected to be zero or very close to zero. For example, in the cases of proportions, the total range is naturally defined as $[0,1]$. Where there is no natural total range, the range can be defined using the ‘10% overshoot rule’ (Cooke, 1991): 10% less than the minimum 5% quantile assessed by any expert, and 10% greater than the maximum of the 95% quantile assessed by any expert.

Evaluation of point estimates

Performance metrics for point estimates focus on assessment of the difference between the prediction r (the expert’s best estimate) and observed value x . For prediction r_n from expert i and outcome x_n , we define the prediction error for quantity X_n as

$$e_n = x_n - r_n, n = 1, \dots, N, \quad (3)$$

where N is the number of quantities being assessed. For a set of predictions, performance is then assessed as the average error across quantities, standardized for differences in question responses scale. Commonly applied indices of accuracy include the Mean Absolute Percentage Error (MAPE), Median Absolute Percentage Error (MdAPE) and Root Mean Absolute Percentage Error (RMAPE) (Hyndman & Koehler, 2006). However, MAPE, RMAPE and related metrics may be overly influenced by performance on individual questions. For instance, if the realized outcome is close to zero, error values are inflated (Makridakis, 1993; Jose & Winkler, 2008). An alternative approach is to standardize each response by the range of responses for that question, known as range-coding.

Expressing each response r_i^n from expert i for quantity n in range-coded form gives

$$r_i^n = \left[\frac{r_i^n - r_{\min}^n}{r_{\max}^n - r_{\min}^n} \right], \quad (4)$$

where r_i^n is the range-coded response, r_{\max}^n is the maximum of the all expert responses assessed for quantity n (including the true value), and r_{\min}^n is the minimum of the all expert responses assessed for quantity n (including the true value). Range-coding is used to equalize the weight or contribution of a variable to the results of an analysis, for example, in character coding in phylogenetic analysis (e.g. Archie, 1985), and raster data in GIS (e.g. de Smith, 2009). Using the range-coded responses, we then calculated performance as the average log-ratio error (ALRE; Burgman *et al.*, 2011),

$$\text{ALRE}_i = \frac{1}{N} \sum_{n=1}^N \left| \log_{10} \left[\frac{x^n + 1}{r_i^n + 1} \right] \right|, \quad (5)$$

where N is the number of quantities, r_i^n is the range-coded prediction, and x^n is the range-coded observed (true) value. Using the range-coded responses, the ALRE has a maximum score of $\log_{10}(2)$ (0.31) when the true value coincides with the group minimum and/or maximum for the question, and a best possible score of zero.

Selection of journal articles

We obtained permission from the editorial board and publisher to use in press and recently available online articles from the journal *Diversity and Distributions*. This is a ‘journal of conservation biogeography’, publishing papers concerned with ‘the application of biogeographical principles, theories and analyses (being those concerned with the distributional dynamics of taxa and assemblages) to problems concerning the conservation of biodiversity’ (Richardson & Whittaker, 2010). It was selected because it addresses a diverse range of ecological systems and contexts, and one of the authors (MAB) was on the editorial board. Four research articles were selected from the journal for use in the study over the course of 2008–2010. Selected articles presented results that could be used to develop prediction questions for experts and covered a broad spectrum of sub-disciplines and study organisms within ecology (Table 1; studies 1–4, respectively). Permission was obtained from the authors to make use of their articles in this research.

Studies 1 and 2 made use of in press articles not available online at the time of the survey. Studies 3 and 4 made use of in press articles available online during the period experts completed the questionnaires. Experts were asked to indicate whether they had previously read a version of the paper. One expert indicated they were familiar with the study in question and their responses were not used in the analysis.

Expert questionnaires

Questionnaires were constructed as interactive pdf forms. Each questionnaire contained a set of demographic and professional

Table 1 Summary of journal articles and expert participants.

Article Synopsis	Areas of expertise addressed	Types of quantities assessed	Number of questions	Number of participants (total number of experts contacted*)	Range of years of relevant experience (median)	Percentage of participants who felt qualified to review the article (%)
1. Range-edge effects on clonal growth in plants (Beatty <i>et al.</i> , 2008)	Plant ecology, population genetics, fragmented landscapes, clonal plants	1. Number of genets per patch 2. Percentage genetic variation between	5	13 (58)	7–35 (14)	85
2. Relationships between habitat use and wombat road fatalities (Roger & Ramp, 2009)	Urban ecology, road ecology, population modelling, mammals	1. Average number of wombat road fatalities 2. Probability of fatality	5	8 (25)	5–16 (15)	44
3. Freshwater fish invasions in mediterranean-climate regions (Marr <i>et al.</i> , 2010)	Invasion biology, aquatic ecology, freshwater fish, mediterranean-climate regions	1. Number of freshwater fish species 2. Percentage of introduced freshwater fish species	13	9 (59)	7–26 (14)	80
4. Bird diversity in urbanizing regions over time (Catterall <i>et al.</i> , 2010)	Urban ecology, avian ecology, Australian birds, South-east Queensland	1. Average number of birds per site 2. Average number of bird species per site	12	12 (67)	2–20 (11)	68

*After returned email addresses were removed.

questions of the experts themselves, and a further 8–13 questions addressing the primary outcomes of the study, hereafter termed the evaluation questions. Each questionnaire provided an overview of the article aims, a summary of any relevant definitions and methods, and information explaining how to use the 4-point estimation method for interval judgments, based on the work of Speirs-Bridge *et al.* (2010). Each questionnaire was designed to be completed within 20 to 30 min (refer to Appendix S1–S4 in Supporting Information for complete questionnaires).

Demographic questions provided information on experts' current position and institutional affiliations, qualifications, gender, years of experience and areas of professional expertise. In addition, experts were asked 'If asked, would you have felt qualified to review this article?' (Y/N). They indicated their level of familiarity with the study on a 4-point scale ranging from 1. not at all familiar (e.g. no prior knowledge) to 4. very familiar (e.g. reviewer, colleague, collaborator, etc.). An additional question was included in studies 3 and 4 that asked experts to rank their expertise for answering the questions on a scale of 1–10 (with 1 being 'no knowledge', and 10 being 'the person most qualified to answer this questionnaire, not including the authors of this study'). This question was designed to assess the ability of experts to rate their own expertise and to provide a more direct assessment of expertise than years experience or number of publications (which relate to seniority).

Evaluation questions

Evaluation questions asked experts to predict several quantities including averages, percentages, probabilities and proportions. Experts were directed to give estimates of an observed experimental result, not any underlying 'true' value. Thus, their responses took into account anticipated errors and biases in results introduced by the experimental methods. Exactly which quantities experts were asked to estimate was determined by the type and number of results in the article. Where possible, quantities were selected for which the expert would have relevant direct experience.

Question format followed a 4-point estimation method that has been shown to reduce overconfidence in expert responses (Speirs-Bridge *et al.*, 2010). Using this method, experts are asked to estimate their (1) Lower limit, (2) Upper limit, (3) Best guess, and (4) Level of confidence that the reported value of the variable in the given study lies between these limits (Fig. 1). A background information document containing the complete introduction and methods sections with the most relevant information highlighted for each article was provided to the experts.

Participants

Experts

Keyword searches were conducted on Google Scholar and ISI Web of Science to identify experts who had published on

1a.) The number of introduced freshwater fish species in California:

i. Realistically, the number of introduced species could be as low as species

ii. Realistically, the number of introduced species could be as high as species

iii. Best guess of the number of introduced species species

iv. For the interval I've created above (from lower to upper), I think the chance that the number of introduced species observed will fall in this interval is: %

(type a number between 0 and 100)

Figure 1 Example question using the 4-point estimation method.

topics closely related to each journal article. Additional experts with relevant experience were identified via referrals from colleagues. A total of 209 experts were contacted, of which 42 experts returned completed questionnaires, a response rate of 20% (Table 1). An additional 15% of experts responded to the email invitation declining to participate either because of time commitments (10%) or because they felt underqualified to complete the questionnaire (5%). Participants were primarily affiliated with universities (67%) or government (24%). The majority (75%) of participants were men. They had a median of 14 years experience, and with the exception of two participants reported 5 years or more relevant experience. An average of 74% of experts who participated felt qualified to review the article they responded to questions on. The average self-rating of expertise on a scale of 1–10 was 6 and 5.75 for studies 3 and 4, respectively.

Students

Responses were solicited for study 3 ('invasive fish') from students in a third-year marine botany course at the University of Melbourne. A total of 17 students completed questionnaires. The majority (62.5%) of student participants were women. Three students were in a Masters program and held Bachelors degrees, and 14 were final year undergraduates. One student also worked as a technician, and one student was affiliated with the private sector and had 4 years experience as a research consultant.

Procedure

Experts

Invitations were emailed to experts on a rolling basis over several weeks for each questionnaire. For studies 1 and 2, additional experts were contacted in a second round of emails to supplement initial response numbers. In total, groups of up to 50 experts were contacted for each study to compensate for low response rates. Selected experts were emailed with information about the study, an invitation to participate, the background information and expert questionnaire. Experts were asked to first review the aims and methods of the study covered in the background documentation and then to complete the questionnaire. They were informed that all responses would remain anonymous. Experts were asked to

complete responses within a month and were sent a reminder email 1 week before the specified due date, although some responses were received much later. For questionnaires completed once the article was available online, we ensured the respondents had not seen the final publication before answering.

Students

Students in the marine biology class were given the opportunity to complete the questionnaire voluntarily for the chance of a book voucher for the best performing responses. They completed the questionnaire during an initial 20-min class session. Students were given a brief introductory presentation describing the aims of the project prior to completing the questionnaire. They received the same background information and completed the same questionnaire as the experts, and answered a subset of the professional experience questions presented to experts, including occupation, affiliation, qualifications, years experience, gender and areas of professional experience.

Analysis of responses

A total of 580 response sets were analysed across 31 questions (Table 1). Expert responses were evaluated against observed experimental results in the four articles. The results reported in each article are themselves subject to sampling and measurement error, and so serve only as an approximation of the truth. Experts were asked to keep this in mind when making their estimates; though, we recognize how to do this is not straightforward. Observations of expert discussions of reasoning processes at a workshop in which herpetologists answered a similar set of evaluation questions suggest that experts attempt to account for the experimental methods in generating their estimates. However, for the purpose of these analyses, we treat the data reported as the 'truth'. For example, when we discuss intervals as capturing the truth, we mean including the values reported in the journal article.

Evaluation of interval calibration and informativeness

The 4-point method frequently results in intervals of different assigned confidence levels. Intervals were standardized to a common confidence level of 80% to allow for comparison. We refer to these intervals as derived 80% intervals. Adjustments were made by fitting a beta-PERT distribution to responses for each question (lower bound, upper bound, best guess and confidence level), a specially constrained beta distribution often used for expert judgments, that is fitted using estimates of minimum, maximum and most likely values (Vose, 1996). Parameters were fitted using least squares regression and treating the best guess as the distribution mode, providing a good approximation of the expert's uncertainty in almost all cases. Interval hit rates are reported using these derived 80% intervals.

To examine differences in calibration at different levels of confidence, unadjusted responses were grouped according to

assigned interval confidence for the experts in each study (i.e. 0.5, 0.6, 0.7, 0.8, 0.9, 1; rounded to the closest probability category). These were then used to generate calibration curves (Lichtenstein & Fischhoff, 1977), plots of expected hit rates against observed hit rates for each study. Interval informativeness was evaluated using Cooke's relative information score (equation 1) for each expert. We undertook a meta-analysis to pool individual study estimates and determine an overall mean hit rate (Borenstein *et al.*, 2009; Harrison, 2010). We used the method of Hedges and Vevea (Hedges & Olkin, 1985; Hedges & Vevea, 1998) and fitted a random-effects model to account for between-study variance, providing inferences beyond studies used in the analysis.

Evaluation of point estimates

Average log-ratio error (ALRE) scores were calculated for each expert's best estimates. We compared the average expert performance, the best and worst performing expert, the most experienced and best ranked expert (according to expert self-assessment) for each study. We took a linear combination of the responses from each expert (i.e. the average) for each question and calculated ALRE scores to assess the benefit of pooling inputs from multiple experts (e.g. see Clemen & Winkler, 1999).

Relationships between performance and expert qualifiers

We tested for pairwise correlations between the three performance metrics (hit rate, ALRE and informativeness) and years experience, number of publications, self-assessment of expertise, and reviewer status for experts in each of the four study expert groups. We applied the meta-analysis approach outlined above to calculate an overall mean correlation for each pairwise relationship.

RESULTS

In all four studies, experts exhibited overconfidence at the 80% level (Table 2). Average hit rates ranged from 49% to 63%, all well below the desired 80% confidence level. Hit rates varied substantially between individuals and questions. Of the 42 expert participants, 16 provided responses for which the

Table 2 Average hit rates with 95% confidence intervals for the expert and student groups.

Experience level	Study		Average hit rate (%)	95% CI
Experts	1	Clonal plants	49	33–66
	2	Wombat fatalities	58	34–82
	3	Invasive fish	57	40–74
	4	Bird diversity	63	48–79
	Pooled		57	51–63
Students	5	Invasive fish – students	76	63–90

derived 80% intervals were well calibrated (defined here as $80 \pm 5\%$). Six experts had derived 80% intervals that captured the truth 100% of the time and two experts gave responses for which the derived intervals had zero hits. The calibration curves for the pooled, unadjusted responses from each expert group revealed similarly high levels of overconfidence (Fig. 2). In each case, probabilities assigned to each probability category (e.g. 0.5, 0.6, 0.7) corresponded poorly or not at all to the observed frequencies.

Student performance

The student group had an average hit rate of 76% (95% CI = 63–90%), exhibiting close to perfect calibration. While they achieved a higher hit rate than the expert groups, it was not significantly different from the corresponding expert group (Table 2). Similar to the expert groups, the inter-student variation in average hit rate was high. Differences in the spread of responses for expert and student groups are illustrated in Fig. 3 for study 3, question 1. In this question, 4 of 9 experts and 12 of 17 students provided 80% intervals capturing the truth, group hit rates of 44% and 71%, respectively. Students achieved higher hit rates by providing much wider intervals. Similar response patterns were observed in questions 2–13.

Point estimates

Average error levels for each expert varied considerably between individual questions (Fig. 4; reflected in the wide error bars). Performance for the most experienced expert was

mixed: they performed worse than average for studies 1, 2 and 3 and corresponded to the best performing expert for study 4. For studies 3 and 4, where participants supplied self-assessments of expertise, the highest ranked experts corresponded in both cases to the expert with the worst overall performance. In all studies, the performance of the responses constructed using the linear opinion pool (average) of the expert responses was better than the performance of the average individual.

Comparison of ALRE scores for expert and student responses in study 3 suggested experts were slightly more accurate on average (Table 3). The average student and student linear opinion pool performed slightly worse than their expert counterparts. However, differences were small and not significant at a 95% level. The best performing student performed better than the average expert, and the worst performing expert was outperformed by the student average.

Relationships between performance and expert status

Correlations between the three performance metrics and between years experience, number of publications, and reviewer status are reported in Table 4. Hit rate was negatively correlated with informativeness in all four studies (overall mean Pearson's $\bar{r} = -0.55$, 95% CI = -0.75 to -0.25) suggesting that experts achieved higher hit rates by providing wider, less informative intervals (Fig. 5). Years experience was positively related to number of publications (Pearson's $\bar{r} = -0.37$, 95% CI = 0.04 – 0.64) and self-assessments were positively related with higher (worse) ALRE scores (Pearson's $\bar{r} = 0.48$, 95% CI = 0.02 – 0.77),

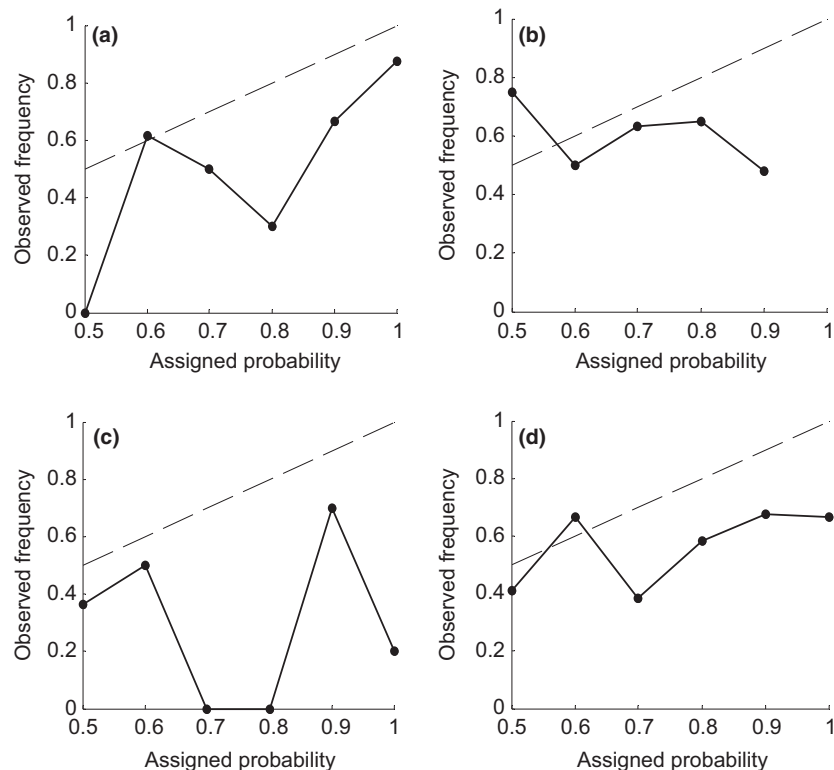


Figure 2 Comparison of assigned probability to observed frequency ('calibration curves') for unadjusted judgments pooled across experts for the (a) clonal plants, (b) wombat fatalities, (c) invasive fish and (d) bird diversity studies. Dashed lines correspond to perfect calibration and solid lines depict the observed expert frequencies. Points on a calibration curve falling below the dashed line signal overconfidence.

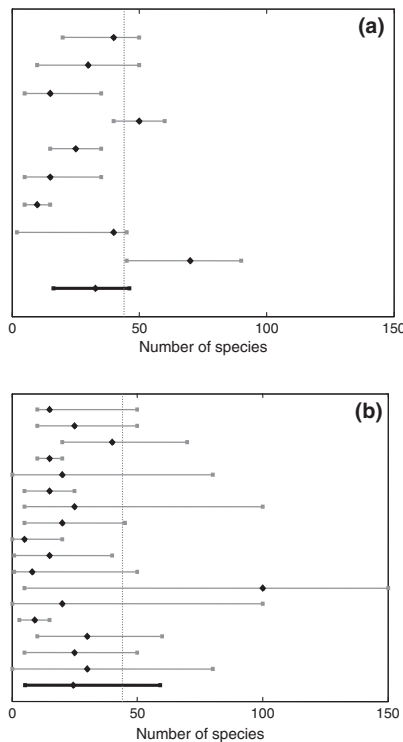


Figure 3 Example of the spread of responses for question 1, study 3 for the (a) experts ($n = 9$), and (b) students ($n = 17$). Individual responses are reported along the y-axes. Grey lines depict the 80% derived intervals for each participant, with black dots marking the location of the best guess. The grey vertical lines show the true value. The linear opinion pool estimates constructed from responses in each group are shown at the bottom in solid black.

suggesting that individuals with greater self-assessed expertise may tend to have larger errors.

Experts with lower (better) ALRE scores tended to achieve higher hit rates (Pearson's $\bar{r} = -0.40$, 95% CI = -0.65 to 0.06) and provide more informative intervals (Pearson's $\bar{r} = -0.19$, 95% CI = -0.50 to 0.16). There was some also evidence that experts with greater numbers of publications had lower hit rates (Pearson's $\bar{r} = -0.25$, 95% CI = -0.55 to 0.10). However, these relationships are tentative and calculated confidence intervals for the pooled correlations may be conservative (i.e. too narrow), as difficulties arise in correctly estimating between-study variance for small numbers of studies (Hedges & Vevea, 1998). The contradictory trends observed across studies for many of the pairings also suggest the presence of additional sources of variance that are unaccounted for. Experts who felt qualified to review the journal articles also had lower hit rates and higher ALRE scores than those who did not feel qualified in all four cases. However, the small sample sizes involved precluded any formal testing of this relationship.

DISCUSSION

Experts had consistently high levels of overconfidence in all four case studies. The derived 80% confidence intervals based

on the experts' responses captured the truth on average only 49–63% of the time. This is in line with the rates of 30–60% overconfidence (i.e. hit rates of 30–60% for 90% intervals) typically reported in the literature (e.g. Önköl *et al.*, 2003; Soll & Klayman, 2004; Speirs-Bridge *et al.*, 2010). What is striking, however, is the contrast between expert and student calibration. While experts were 17–31% overconfident, students' derived 80% confidence intervals captured the truth on average 76% of the time, meaning they were, on average, almost perfectly calibrated. Differences in hit rates between experts and between experts and students could largely be accounted for by the precision of the intervals they provided (Table 4, row 2; Figs 3 & 5); few individuals excelled at both. More knowledgeable individuals frequently achieve similar levels of calibration to less knowledgeable individuals by providing better placed, but narrower intervals (e.g. Önköl *et al.*, 2003; McKenzie *et al.*, 2008). This accuracy-precision trade-off (Yaniv & Foster, 1995; Yaniv, 1997) has been attributed to social pressure to provide informative predictions: people value informative, inaccurate forecasts over accurate but less informative forecasts (Yates *et al.*, 1996). Students may have provided appropriately wide intervals because they did not consider themselves to be expert.

The consistent finding of overconfidence has important implications for the use of expert estimates of uncertainty. If responses from this study had been used as priors in a Bayesian analysis for example, they would have been overly precise and misleading given the experts' knowledge. If results from these studies generalize to other cases, they suggest that bounds provided in studies such as O'Neill *et al.* (2008) and Murray *et al.* (2009) might be overconfident. As Fig. 5 illustrates, the trade-off between the interval accuracy and precision is marked. Results from other fields suggest that training in the form of practice with adequate feedback may reduce overconfidence (Bolger & Önköl-Atay, 2004). Even two or three feedback sessions can assist in improving calibration levels for poorly calibrated experts within their domain (e.g. Lichtenstein *et al.*, 1982). We speculate that if experts were systematically exposed to such feedback as part of their formal training, assessments would be much better calibrated.

Previous studies of expert judgment have consistently found no relationship between performance and metrics such as years experience or peer status (e.g. Armstrong, 1980; Camerer & Johnson, 1991; Shanteau, 1992; Cooke *et al.*, 2008; McKenzie *et al.*, 2008; Burgman *et al.*, 2011). Our findings support this. Interestingly, we also found evidence that individuals with greater self-assessments of expertise had less accurate ALRE scores. A recent study involving a similar style of expertise self-assessment noted this same relationship (Önköl *et al.*, 2009). Several authors have speculated that greater expertise may reduce performance by leaving individuals more susceptible to confirmation bias (e.g. Armstrong, 1980; Camerer & Johnson, 1991; Koehler, 1993). They become less likely to seek counterfactual evidence and thus may be outperformed by non-experts who use such evidence. Tetlock (2005), for example, found that style of reasoning rather than level of

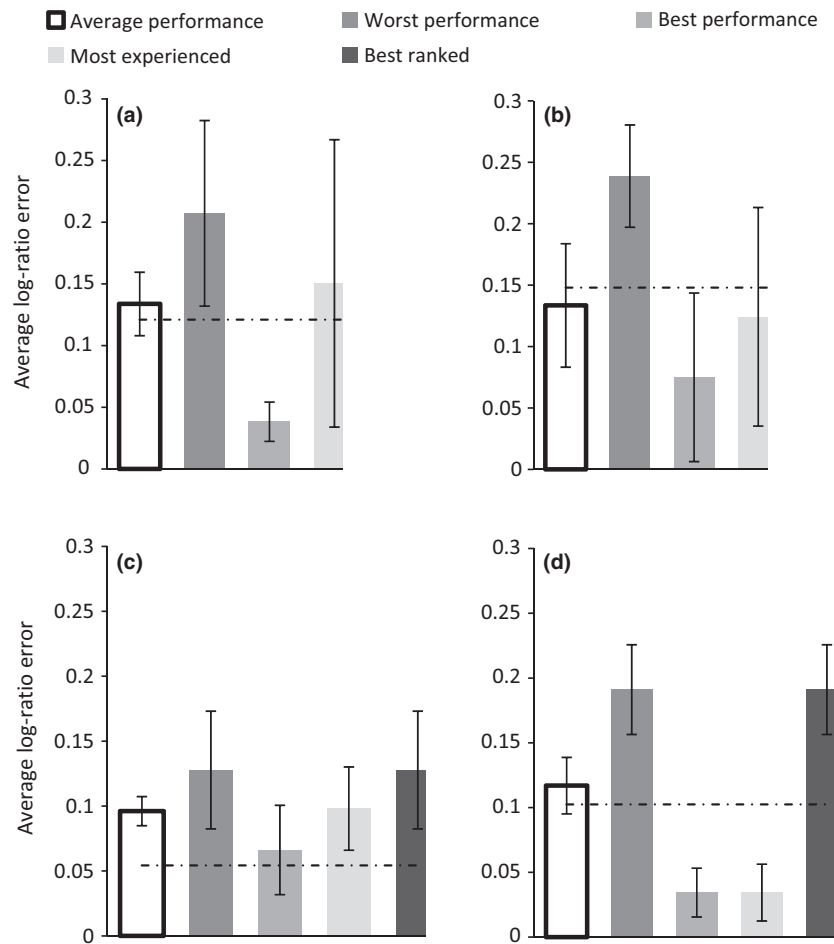


Figure 4 Summary of average log-ratio error (ALRE) scores for the (a) clonal plants, (b) wombat fatalities, (c) invasive fish and (d) bird diversity studies. Lower ALRE scores correspond to greater accuracy. Error bars for average performance (of group members) represent 95% confidence intervals (CI) across group members. Error bars in all other cases represent 95% CIs on ALRE scores across questions. Dashed lines show performance of the group linear opinion pool. Note that ALRE scores are scaled according to the set of responses for a particular expert group and are not directly comparable across studies.

Table 3 Average log-ratio error (ALRE) scores for study 3 ('Invasive fish') for the expert and student groups. Lower ALRE scores correspond to greater accuracy.

	Experts	Students
Average performance	0.07 (0.06–0.08)	0.08 (0.07–0.10)
Worst performance	0.10 (0.06–0.13)	0.12 (0.08–0.16)
Best performance	0.05 (0.03–0.07)	0.06 (0.04–0.08)
Most experienced	0.07 (0.05–0.10)	
Best ranked	0.10 (0.06–0.13)	
Opinion pool	0.04 (0.02–0.06)	0.05 (0.03–0.07)

Values in parentheses show 95% confidence intervals.

expertise predicted performance in forecasting geo-political events. The individuals who performed best were those with a more 'open' mindset, willing to contemplate multiple possible scenarios and hypotheses. The implications for decision-makers are to avoid the temptation to rely on input from one or a few specialist experts. In the absence of explicit information on performance, aggregate measures such as the linear opinion pool are likely to represent the best course of action for decision-makers (Clemen & Winkler, 1999; Arm-

strong, 2001). The robust performance of the group opinion pool observed here is consistent with this recommendation.

Study limitations

Evaluation studies that report poor expert performance are sometimes criticized because the tasks were too hard or not appropriate, and/or the experts were not the right experts (e.g. Dawes *et al.*, 1989; Tetlock, 2005). Ours was an admittedly artificial context in the sense that the experts do not necessarily represent the group of experts that would actually have been convened to answer these questions in a real expert judgment context. However, the experts had all worked or published directly on the topic of interest or on a closely related topic. While the questions may be considered to be 'hard' in the sense that experts were requested to make inferences outside their direct field of expertise and experience, we would argue that they are representative of the type and level of difficulty often required by experts.

Further work remains to see whether and how expert performance differs across different task environments within ecology. In our analyses, small sample sizes may have precluded detection of trends between performance and experience, publications and self-assessment of expertise.

Table 4 Pearson correlations between performance measures (hit rate, ALRE, informativeness) and expert qualifiers (years experience, number of publications and self-assessment of expertise) for experts in each study and pooled across studies.

Study	Correlation	1	2	3	4	Pooled mean* [95% CI]
		Clonal plants	Wombat fatalities	Invasive fish	Bird diversity	
Hit rate	ALRE	−0.08 (0.81)	−0.63 (0.10)	−0.69 (0.04)	−0.33 (0.30)	−0.40 [−0.65 to 0.06]
	Informativeness	−0.59 (0.03)	−0.11 (0.81)	−0.58 (0.10)	−0.66 (0.02)	−0.55 [−0.75 to −0.25]
	Experience	−0.20 (0.51)	−0.03 (0.95)	0.42 (0.26)	0.18 (0.58)	0.07 [−0.28 to 0.40]
	Publications	−0.16 (0.59)	−0.35 (0.40)	−0.38 (0.32)	−0.20 (0.54)	−0.25 [−0.55 to 0.10]
	Self-assessment			−0.27 (0.48)	−0.07 (0.84)	−0.15 [−0.58 to 0.34]
ALRE	Informativeness	−0.03 (0.93)	−0.46 (0.26)	0.05 (0.90)	−0.36 (0.26)	−0.19 [−0.50 to 0.16]
	Experience	−0.31 (0.30)	0.35 (0.40)	0.13 (0.73)	−0.57 (0.05)	−0.19 [−0.55 to 0.23]
	Publications	−0.47 (0.10)	0.14 (0.75)	0.74 (0.02)	−0.15 (0.65)	0.07 [−0.49 to 0.59]
	Self-assessment			0.65 (0.06)	0.34 (0.28)	0.48 [0.02 to 0.77]
Informativeness	Experience	0.62 (0.03)	−0.23 (0.59)	−0.25 (0.51)	0.07 (0.82)	0.13 [−0.33 to 0.54]
	Publications	0.34 (0.25)	0.23 (0.58)	−0.31 (0.41)	0.12 (0.72)	0.13 [−0.23 to 0.45]
	Expertise			−0.50 (0.17)	−0.02 (0.96)	−0.23 [−0.63 to 0.27]
Experience	Publications	0.44 (0.13)	0.67 (0.07)	0.04 (0.93)	0.30 (0.35)	0.37 [0.04 to 0.64]
	Self-assessment			0.04 (0.92)	−0.03 (0.92)	0.00 [−0.47 to 0.47]
Self-assessment	Publications			0.94 (0.00)	−0.30 (0.35)	0.60 [−0.86 to 0.99]

ALRE, average log-ratio error.

Correlations between variables > 0.5 are highlighted in grey.

*Pooled mean estimates for correlations involving self-assessment are calculated on the basis of two studies. Random-effects models perform poorly for small study numbers and, particularly where between-group heterogeneity is high, these bounds are likely to be overly conservative (i.e. too narrow).

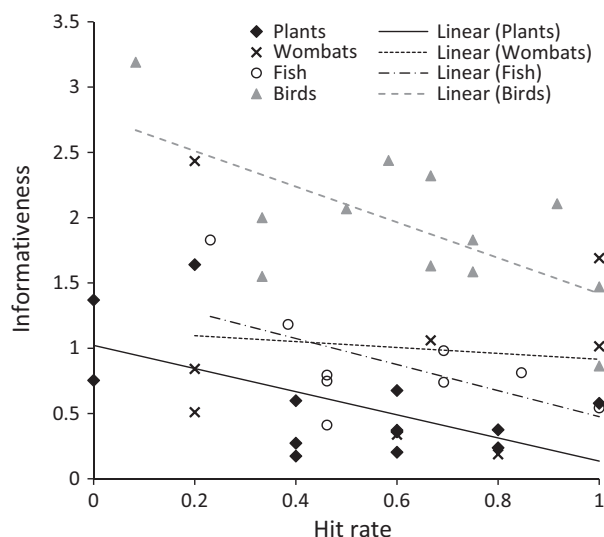


Figure 5 Relationship between hit rate and informativeness for experts in each study. Least squares regression lines are displayed for each expert group. The negative slopes illustrate the trade-off between accuracy (hit rate) and precision (informativeness) for experts in specifying interval responses. Informativeness scores are specific to the questions asked and are not directly comparable across studies.

Additional studies of expert judgment performance would test our tentative conclusions. More refined, task-specific measures of expertise than years experience and number of publications could be tested; however, the literature overwhelmingly does

not support the idea that this will improve our ability to predict performance.

Future directions

The results of this study suggest that it is important to evaluate expert opinion in ecology. Evaluation allows us to learn more about expert knowledge, its limitations and when it can be applied reliably. Experts possess valuable knowledge, but may require training to convey this knowledge accurately to decision-makers. In the absence of training or information about past performance, use of aggregate responses may represent the most reliable approach, providing input from multiple experts and countering individual variation in performance.

ACKNOWLEDGEMENTS

The authors would especially like to thank each of the experts who kindly participated in this study for their time and interest. We thank Gemma Beatty, Jim Provan, Erin Roger, Daniel Rank, Sean Marr, Carla Catterall and Jarrad Cousin and their associated co-authors for permission to use their articles and results as a part of this study. We are thankful for assistance and permissions from David Richardson and Josephine de Mink, Jan Carey, and helpful discussions and ideas from Andrew Speirs-Bridge, Louisa Flander, Andrew Robinson, Terry Walshe, Bonnie Wintle and Anna Carr. Comments from David Keith and two anonymous reviewers greatly improved this manuscript.

REFERENCES

- Archie, J.W. (1985) Methods for coding variable morphological features for numerical taxonomic analysis. *Systematic Biology*, **34**, 326–345.
- Armstrong, J. (1980) The seer-sucker theory: the value of experts in forecasting. *Technology Review*, **82**, 16–24.
- Armstrong, J.S. (2001) Combining forecasts. *Principles of forecasting: a handbook for researcher and practitioners* (ed. by J.S. Armstrong), pp. 417–440. Kluwer Academic Publishers, Norwell.
- Ashton, R.H. (1974) An experimental study of internal control judgments. *Journal of Accounting Research*, **12**, 143–157.
- Beatty, G.E., McEvoy, P.M., Sweeney, O. & Provan, J. (2008) Range-edge effects promote clonal growth in peripheral populations of the one-sided wintergreen *Orthilia secunda*. *Diversity and Distributions*, **14**, 546–555.
- Bolger, F. (1995) Cognitive expertise research and knowledge engineering. *The Knowledge Engineering Review*, **10**, 3–19.
- Bolger, F. & Önköl-Atay, D. (2004) The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, **20**, 29–39.
- Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. (2009) *Introduction to meta-analysis*. Wiley, Chichester.
- Burgman, M.A. (2005) *Risk and decisions for conservation and environmental management*. Cambridge University Press, Cambridge.
- Burgman, M.A., Lindenmayer, D.B. & Elith, J. (2005) Managing landscapes for conservation under uncertainty. *Ecology*, **86**, 2007–2017.
- Burgman, M.A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. & Twardy, C. (2011) Expert status and performance. *PLoS ONE*, **6**, e22998.
- Camerer, C.F. & Johnson, E.J. (1991) The process-performance paradox in expert judgment: how can experts know so much and predict so badly? *Towards a general theory of expertise: prospects and limits* (ed. by K.A. Ericsson and J. Smith), pp. 195–217. Cambridge Press, New York.
- Campbell, L.M. (2002) Science and sustainable use: views of marine turtle conservation experts. *Ecological Applications*, **12**, 1229–1246.
- Carpenter, S.R. (2002) Ecological futures: building an ecology of the long now. *Ecology*, **83**, 2069–2083.
- Catterall, C.P., Cousin, J.A., Piper, S. & Johnson, G. (2010) Long-term dynamics of bird diversity in forest and suburb: decay, turnover or homogenization? *Diversity and Distributions*, **16**, 559–570.
- Christensen-Szalanski, J.J.J., Diehr, P.H. & Bushyhead, J.B. (1982) Two studies of good clinical judgment. *Medical Decision Making*, **2**, 275–283.
- Clemen, R.T. & Winkler, R.L. (1999) Combining probability distributions from experts in risk analysis. *Risk Analysis*, **19**, 187–203.
- Cooke, R.M. (1991) *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, New York.
- Cooke, R.M., ElSaadany, S. & Huang, X.Z. (2008) On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering and System Safety*, **93**, 745–756.
- Crome, F.H.J., Thomas, M.R. & Moore, L.A. (1996) A novel Bayesian approach to assessing impacts of rain forest logging. *Ecological Applications*, **6**, 1104–1123.
- Czembor, C.A. & Vesk, P.A. (2009) Incorporating between-expert uncertainty into state-and-transition simulation models for forest restoration. *Forest Ecology and Management*, **259**, 165–175.
- Dawes, R.M., Faust, D. & Meehl, P.E. (1989) Clinical versus actuarial judgment. *Science*, **243**, 1668–1674.
- Doswald, N., Zimmermann, F. & Breitenmoser, U. (2007) Testing expert groups for a habitat suitability model for the lynx *Lynx lynx* in the Swiss Alps. *Wildlife Biology*, **13**, 430–446.
- Ericsson, K.A. (2004) Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, **79**, S70–S81.
- Ericsson, K.A. & Kintsch, W. (1995) Long-term working memory. *Psychological Review*, **102**, 211–245.
- Ericsson, K.A. & Lehmann, A.C. (1996) Expert and exceptional performance: evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, **47**, 273–305.
- Fazey, I., Fazey, J.A. & Fazey, D.M.A. (2005) Learning more effectively from experience. *Ecology and Society*, **10**, 4. Available at: <http://www.ecologyandsociety.org/vol10/iss2/art4/> (accessed 16 April 2008).
- Fischhoff, B. (1990) Understanding long-term environmental risks. *Journal of Risk and Uncertainty*, **3**, 315–330.
- Harrison, F. (2010) Getting started with meta analysis. *Methods in Ecology and Evolution*, **2**, 1–10.
- Hearst, E. (1988) Fundamentals of learning and conditioning. *Steven's handbook of experimental psychology* (ed. by R.C. Atkinson, R.J. Herrnstein, G. Lindzey and R.D. Luce), pp. 1–109. Wiley-Interscience, New York.
- Hedges, L.V. & Olkin, I. (1985) *Statistical methods for meta-analysis*. Academic Press, New York.
- Hedges, L.V. & Vevea, J.L. (1998) Fixed- and random-effects models in meta-analysis. *Psychological Methods*, **3**, 486–504.
- Hilborn, R. & Ludwig, D. (1993) The limits of applied ecological research. *Ecological Applications*, **3**, 550–552.
- Hyndman, R.J. & Koehler, A.B. (2006) Another look at measures of forecast accuracy. *International Journal of Forecasting*, **22**, 679–688.
- Iglesias, R.M.R. & Kothmann, M.M. (1998) Evaluating expert knowledge: plant species responses to cattle grazing and fire. *Journal of Range Management*, **51**, 332–344.
- Irvine, R., Fiorini, S., Yearley, S., McLeod, J., Turner, A., Armstrong, H., White, P. & Van Der Wal, R. (2009) Can managers inform models? Integrating local knowledge into models of red deer habitat use. *Journal of Applied Ecology*, **46**, 344–352.
- Johnson, C.J. & Gillingham, M.P. (2004) Mapping uncertainty: sensitivity of wildlife habitat ratings to expert opinion. *Journal of Applied Ecology*, **41**, 1032–1041.

- Jorgensen, M., Teigen, K.H. & Molokken, K. (2004) Better sure than safe? Over-confidence in judgement based software development effort prediction intervals. *The Journal of Systems and Software*, **70**, 79–93.
- Jose, V.R.R. & Winkler, R.L. (2008) Simple robust averages of forecasts: some empirical results. *International Journal of Forecasting*, **24**, 163–169.
- Kaplan, S. (1992) 'Expert information' versus 'expert opinions'. Another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Reliability Engineering and System Safety*, **35**, 61–72.
- Keren, G. (1987) Facing uncertainty in the game of bridge: a calibration study. *Organizational Behavior and Human Decision Processes*, **39**, 98–114.
- Koehler, J.J. (1993) The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, **56**, 28–55.
- Kuhnert, P.M., Martin, T.G. & Griffiths, S.P. (2010) A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecology Letters*, **13**, 900–914.
- Lele, S.R. & Allen, K.L. (2006) On using expert opinion in ecological analyses: a frequentist approach. *Environmetrics*, **17**, 683–704.
- Lichtenstein, S. & Fischhoff, B. (1977) Do those who know more also know more about how much they know. *Organizational Behavior and Human Performance*, **20**, 159–183.
- Lichtenstein, S. & Fischhoff, B. (1980) Training for calibration. *Organizational Behavior and Human Performance*, **26**, 149–171.
- Lichtenstein, S., Fischhoff, B. & Phillips, L.D. (1982) Calibration of probabilities: the state of the art to 1980. *Judgment under uncertainty: heuristics and biases* (ed. by D. Kahneman and A. Tversky), pp. 306–334. Cambridge University Press, Cambridge.
- Makridakis, S. (1993) Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, **9**, 527–529.
- Marr, S.M., Marchetti, M.P., Olden, J.D., Garcia-Berthou, E., Morgan, D.L., Arismendi, I., Day, J.A., Griffiths, C.L. & Skelton, P.H. (2010) Freshwater fish introductions in mediterranean-climate regions: are there commonalities in the conservation problem? *Diversity and Distributions*, **16**, 606–619.
- McCarthy, M.A., Keith, D., Tietjen, J., Burgman, M.A., Maunder, M., Master, L., Brook, B.W., Mace, G., Possingham, H.P., Medellin, R., Andelman, S., Regan, H., Regan, T. & Ruckelshaus, M. (2004) Comparing predictions of extinction risk using models and subjective judgement. *Acta Oecologica – International Journal of Ecology*, **26**, 67–74.
- McCoy, E.D., Sutton, P.E. & Mushinsky, H.R. (1999) The role of guesswork in conserving the threatened sand skink. *Conservation Biology*, **13**, 190–194.
- McKenzie, C.R.M., Liersch, M.J. & Yaniv, I. (2008) Overconfidence in interval estimates: what does expertise buy you? *Organizational Behavior and Human Decision Processes*, **107**, 179–191.
- Morgan, M.G., Pitelka, L.F. & Shevliakova, E. (2001) Elicitation of expert judgments of climate change impacts on forest ecosystems. *Climatic Change*, **49**, 279–307.
- Murphy, A.H. (1993) What is a good forecast: an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–293.
- Murphy, A.H. & Winkler, R.L. (1977) Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, **26**, 41–47.
- Murphy, A.H. & Winkler, R.L. (1984) Probability forecasting in meteorology. *Journal of the American Statistical Association*, **79**, 489–500.
- Murray, J.V., Goldizen, A.W., O'Leary, R.A., McAlpine, C.A., Possingham, H.P. & Choy, S.L. (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies *Petrogale penicillata*. *Journal of Applied Ecology*, **46**, 842–851.
- O'Neill, S.J., Osborn, T.J., Hulme, M., Lorenzoni, I. & Watkinson, A.R. (2008) Using expert knowledge to assess uncertainties in future polar bear populations under climate change. *Journal of Applied Ecology*, **45**, 1649–1659.
- Önkal, D., Yates, J.F. & Simga-Mugan, C. (2003) Professional vs. amateur judgment accuracy: the case of foreign exchange rates. *Organizational Behavior and Human Decision Processes*, **91**, 169–185.
- Önkal, D., Goodwin, P., Thomson, M., Gonul, S. & Pollock, A. (2009) The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, **22**, 390–409.
- Oskamp, S. (1965) Overconfidence in case-study judgments. *Journal of Consulting Psychology*, **29**, 261–265.
- Pearce, J.L., Cherry, K., Drielsma, M., Ferrier, S. & Whish, G. (2001) Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412–424.
- Richardson, D.M. & Whittaker, R.J. (2010) Conservation biogeography: foundations, concepts and challenges. *Diversity and Distributions*, **16**, 313–320.
- Roger, E. & Ramp, D. (2009) Incorporating habitat use in models of fauna fatalities on roads. *Diversity and Distributions*, **15**, 222–231.
- Rothlisberger, J.D., Lodge, D.M., Cooke, R.M. & Finnoff, D.C. (2010) Future declines of the binational Laurentian Great Lakes fisheries: the importance of environmental and cultural change. *Frontiers in Ecology and the Environment*, **8**, 239–244.
- Roura-Pascual, N., Richardson, D.M., Krug, R.M., Brown, A., Chapman, R.A., Forsyth, G.G., Le Maitre, D.C. & Robertson, M.P. (2009) Ecology and management of alien plant invasions in South African fynbos: accommodating key complexities in objective decision making. *Biological Conservation*, **142**, 1595–1604.
- Savage, L.J. (1954) *The foundations of statistics*. Wiley, New York.

- Scholes, R.J. & Biggs, R. (2006) A biodiversity intactness index. *Nature*, **434**, 45–49.
- Seoane, J., Bustamante, J. & Diaz-Delgado, R. (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512–522.
- Shanteau, J. (1992) Competence in experts: the role of task characteristics. *Organizational Behavior and Human Decision Processes*, **53**, 252–266.
- de Smith, M. (2009) *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador Ltd, Leicester.
- Soll, J.B. & Klayman, J. (2004) Overconfidence in interval estimates. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **30**, 299–314.
- Speirs-Bridge, A., Fidler, F., McBride, M.F., Flander, L., Cumming, G. & Burgman, M. (2010) Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, **30**, 512–523.
- Stanley, T.R. & Skagen, S.K. (2007) Estimating the breeding population of long-billed curlew in the United States. *The Journal of Wildlife Management*, **71**, 2556–2564.
- Sutherland, W.J. (2006) Predicting the ecological consequences of environmental change: a review of the methods. *Journal of Applied Ecology*, **43**, 599–616.
- Tetlock, P.E. (2005) *Expert political judgment: how good is it? How can we know?*. Princeton University Press, Princeton.
- Vose, D. (1996) *Quantitative risk analysis: a guide to Monte Carlo simulation modelling*. Wiley, Chichester.
- Weiss, D.J., Shanteau, J. & Harries, P. (2006) People who judge people. *Journal of Behavioral Decision Making*, **19**, 441–454.
- Whitfield, D.P., Ruddock, M. & Bullman, R. (2008) Expert opinion as a tool for quantifying bird tolerance to human disturbance. *Biological Conservation*, **141**, 2708–2717.
- Winkler, R.L. (1967) The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, **662**, 1105–1120.
- Yamada, K., Elith, J., McCarthy, M. & Zenger, A. (2003) Eliciting and integrating expert knowledge for wildlife habitat modelling. *Ecological Modelling*, **165**, 251–264.
- Yaniv, I. (1997) Weighting and trimming: heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, **69**, 237–249.
- Yaniv, I. & Foster, D.P. (1995) Graininess of judgment under uncertainty: an accuracy-informativeness trade-off. *Journal of Experimental Psychology. General*, **124**, 424–432.
- Yates, J.F., Price, P.C., Lee, J.W. & Ramirez, J. (1996) Good probabilistic forecasters: the ‘consumers’ perspective. *International Journal of Forecasting*, **12**, 41–56.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Expert questionnaire for study 1.

Appendix S2 Expert questionnaire for study 2.

Appendix S3 Expert questionnaire for study 3.

Appendix S4 Expert questionnaire for study 4.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCHES

Marissa F. McBride is a doctoral student in the School of Botany at the University of Melbourne. Her dissertation research focuses on methods for evaluating and improving the use of expert knowledge in environmental decision-making.

Fiona Fidler is a Senior Research Fellow at the Australian Centre of Excellence for Risk Analysis. Her current research interests are expert decision-making and statistical cognition. She has an undergraduate degree in cognitive psychology and PhD in philosophy of science.

Mark Burgman is Director of the Australian Centre of Excellence for Risk Analysis and the Adrienne Clarke Chair of Botany at the University of Melbourne. He works on ecological modelling, conservation biology and risk assessment. He received a BSc from the University of New South Wales (1974) and a PhD from the State University of New York (1987). He worked as a consultant ecologist and research scientist in Australia, the United States and Switzerland during the 1980s before joining the University of Melbourne in 1990. He has published five authored books, two edited books, over 150 research papers and more than 50 reviewed reports and commentaries.

Author contributions: all authors conceived the ideas. M.F.M. collected and analysed the data and led the writing under supervision from F.F. and M.A.B. All authors discussed the results and made substantial contributions to the revisions.

Editor: David Richardson