Expert Elicitation: Using the Classical Model to Validate Experts' Judgments

Abigail R. Colson* and Roger M. Cooke†

Introduction

Existing data and modeling tools cannot provide decision makers with all of the information they need to design and implement effective policies and make optimal management choices. Thus decision makers often supplement other forms of information with the judgment of experts. As noted by Morgan and Henrion (1990), if traditional science and statistics cannot provide all of the inputs needed for a model or policy analysis, decision makers have few alternatives to asking experts. Incorporating expert judgment is a way to quantify the uncertainty about otherwise unknown parameters, and it can include methods as disparate as asking a single expert for his or her best guess, informally surveying colleagues, or following a formalized, documented procedure for obtaining and combining probabilistic judgments. The third type of method is called *expert elicitation.*

In the first attempt to standardize the use of expert judgment, the U.S. Nuclear Regulatory Commission (NRC) documented the elicitation process and opened it to scientific review (U.S. Nuclear Regulatory Commission 1975). The NRC uncovered big differences in expert opinion, raising questions about how to validate and combine information gathered from experts (Oppenheimer, Little, and Cooke 2016). Since then, the nuclear safety community has pioneered expert elicitation methods that address the challenges of validating and combining expert judgments, and the techniques have spread to other fields (Cooke 2012), most recently to assess future economic growth (Puig and Morales-Nápoles 2017).

In order to be accepted as scientific data, expert judgments must adhere to the principles of the scientific process, including accountability, neutrality, fairness, and the ability for empirical control (Cooke 1991), the last of which is possible through validation. Validation of expert judgments means both that the judgments reflect the beliefs of the expert and that those beliefs reflect reality. O'Hagan et al. (2006) observe that it is not possible to measure the former, as the only way of knowing the "true" beliefs of experts is through elicitation itself. Validating that beliefs reflect reality, however, can be measured by comparing elicited judgments to observed data where possible.

*University of Strathclyde, Department of Management Science, 199 Cathedral Street, Glasgow G4 0QU, United Kingdom. Tel: +44 (0)141-548-2662; e-mail: abigail.colson@strath.ac.uk.
†Resources for the Future, 1616 P Street, NW, Suite 600, Washington, DC 20036, Tel: 202-328-5000; Fax: 202-939-3460; e-mail: cooke@rff.org.

Review of Environmental Economics and Policy, pp. 1–21
doi: 10.1093/reep/rex022

Unfortunately, expert judgments are rarely validated by observing the variables of interest. Modelers and analysts turn to experts when probabilities of interest are unknown, thus making it difficult to validate the experts' judgments against empirical data. Instead, pundits, blue ribbon panels, and high-level committees ask audiences to trust their judgments based on the credentials of the experts rather than on the experts' track records of making accurate predictions.[1]

Cooke, Mendel, and Thijs (1988) first proposed an expert judgment technique that scores the experts according to their performance against empirical data. The use of these performance scores for creating and validating combinations of expert judgments is called the "classical model," or *structured* expert judgment (Cooke 1991). The classical model has been deployed extensively in a range of areas, including investment banking, volcanology, public health, ecology, and aeronautics. In 2017, a joint report of the National Academies of Science, Engineering, and Medicine recommended the approach for applications used to evaluate the social cost of carbon (National Academies of Sciences, Engineering, and Medicine 2017).

This article, which is part of a symposium on expert elicitation,[2] examines the classical model of structured expert judgment. In the next section we provide a brief introduction to expert elicitation. Then we describe the classical model in more detail and review some alternative approaches for validating and combining experts' judgments. Next, we present a case study of the use of the classical model to inform risk management of invasive species in the U.S. Great Lakes. This is followed by a description of thirty-three applications of the classical model conducted from 2006 through March 2015 and an analysis of the performance of the experts and different schemes for combining and validating expert judgments. In the penultimate section we discuss recent developments regarding the out-of-sample validity of the classical model. We present a summary and conclusions in the final section.

Introduction to Expert Elicitation

Expert elicitation is the process of obtaining probabilistic belief statements from experts about unknown quantities or parameters.[3] Elicited probabilities can supplement other types of evidence and serve as inputs to economic, decision analytic, and other modeling. Elicitation involves carefully defining the target questions, writing and pilot testing an elicitation protocol, training the experts in subjective probability, conducting the interviews, providing feedback to the experts, and analyzing and documenting the results. Major reviews of best practice for eliciting and using expert judgments include Morgan and Henrion (1990), Cooke (1991), and O'Hagan et al. (2006), all of which discuss the need to carefully structure

[1] In some cases, a rigorous and transparent nomination process has been documented to demonstrate that the term "expert" has been appropriately applied (Aspinall and Cooke 2013; Bamber and Aspinall 2013). Although the University of Pennsylvania's recent Good Judgment Project, a large-scale geopolitical event forecasting competition, included some validation of the judgments of participants over a several-year period (Mellers et al. 2015), only a small minority of practitioners perform validation.

[2] The other article is Verdolini et al. (2018), which discusses insights from expert elicitations concerning the prospects for energy technologies.

[3] Experts can also provide decision makers with preferences and other qualitative information. However, our focus here will be on using experts for quantitative inputs.

elicitations to properly capture experts' beliefs and the value of having experts express their beliefs as probabilities.[4]

Expert elicitations typically include multiple experts to capture diversity of knowledge, background, and opinion, but the subsequent modeling or decision problem often requires a single probability distribution for a parameter rather than a set of distributions from several experts. Thus, combining elicited judgments is important. Clemen and Winkler (1999, 2007) review available methods, which are classified as *behavioral*—involving group interaction to arrive at a single consensus distribution—or *mathematical*—using analytic processes on the individual expert assessments to yield one combined distribution, without expert interaction. A common example of a behavioral approach is the Delphi method, which involves multiple rounds of experts providing their assessments and reasoning, sharing that information with all of the experts, and then allowing the experts to revise their assessments, hopefully moving towards consensus (Rowe and Wright 1999).[5] Mathematical techniques for aggregating judgments include axiomatic and Bayesian approaches. Axiomatic methods use simple combination rules to produce a single distribution; examples include the linear and logarithmic opinion pools. Bayesian methods are based on likelihood functions. Behavioral methods can fail to overcome troublesome group dynamics (Cooke 1991). Bayesian methods are difficult to use in practice, while axiomatic approaches are easier to understand and implement (Clemen and Winkler 1999). We focus here on axiomatic combination approaches.

Expert judgment is not an appropriate tool for every quantitative question. For example, Cooke and Goossens (2008) note that expert judgment is not needed when a quantity is observable, such as the speed of light in a vacuum, and not appropriate when a field does not have relevant scientific expertise and related measurements, such as asking about the behavior of a god. They argue that the ideal target for expert judgment is an issue such as the toxicity of a new substance in humans, which is measurable in theory but not in practice.[6] Expert judgment is also not needed if there are sufficient historical data and consensus about the processes for translating that historical data into predictions (Hora 2007). Finally, if an outcome is highly dependent on behavior, predictive expertise may not exist (Morgan 2014).

Eliciting expert judgments is an involved process, requiring time and effort from both the analyst(s) and experts. Quantifying uncertainty may not be worthwhile if it has little impact on the end decision or outcome (Hora 2007). Thus analysts should identify the key potential uncertainties in a problem area prior to conducting an expert judgment study.

Introduction to the Classical Model

The classical model is an approach for eliciting and mathematically aggregating expert judgments, with validation incorporated as a core feature. Experts quantify their uncertainty for two types of questions: *target* questions and *calibration* questions. The variables of interest are the target questions, that is, those that cannot be adequately answered with other methods

[4] Cooke (2015) reviews nonprobabilistic approaches.

[5] If consensus is not reached after a number of rounds, mathematical aggregation can then be used to produce a single assessment.

[6] Related measurements—like the substance's toxicity in other organisms—may exist, and these could be a starting point for expert judgments.

and thus require expert judgment. Experts also assess a set of calibration questions, which are items from the experts' field that are uncertain to the experts (i.e., the experts do not know the true values or do not have the true values readily accessible), but are either known to the analysts at the time of the elicitation or will be known during the analysis period. Experts are scored based on their performance on the calibration questions, and their assessments are weighted (according to their scores) and combined. This performance-based combination is also scored on the calibration questions. Thus the classical model validates both individual expert assessments and the performance-based combinations against observed data.

In the remainder of this section we describe the classical model's approach to eliciting and validating expert judgments and its scoring mechanisms. We also discuss calibration questions in more detail.

Eliciting and Validating Judgments

In the classical model, each expert quantifies his or her uncertainty for each calibration question and variable of interest. This uncertainty quantification could take many forms, but to ensure comparability over a range of applications, the classical model imposes a common structure: experts typically state their fifth, fiftieth, and ninety-fifth percentiles for the estimate of each uncertain item (Cooke 1991; Cooke and Goossens 2008).[7] The fiftieth percentile is the median estimate; the expert thinks it is equally likely that the true value for that item falls above or below the provided value. The fifth and ninety-fifth percentiles create a ninety percent credible range—the expert believes there is a ninety percent chance that the true value falls between those bounds. Sometimes the twenty-fifth and seventy-fifth percentiles are elicited as well, forming a fifty percent credible range (Colson and Cooke 2017).

By providing values for specific percentiles, each expert provides a statistical hypothesis. She says in effect that there is a five percent chance that the true value of the quantity in question falls beneath the fifth percentile, a fifty percent chance the true value falls beneath the fiftieth percentile, and a ninety-five percent chance that the true value falls beneath the ninety-fifth percentile. If she provides assessments for several independent items for which analysts have actual values, then an analyst can observe how frequently the true values fall in the expert's different interpercentile intervals. This provides a mechanism for validation.

To illustrate, if an expert assesses fifth and ninety-fifth percentiles for ten items, she should expect that one of the ten actual values will fall outside the provided interval. Suppose that three realizations fall outside this interval. How poor are the expert's assessments? Assuming the expert's probability statements are correct (i.e., there is a ninety percent chance the true value falls within the ranges given by the expert), the probability of seeing three or more realizations fall outside the ninety percent credible range is 0.07. This means that if we "rejected" this expert, we would have a seven percent chance of rejecting a good expert who had bad luck on these items; 0.07 is the "P-value" at which we would falsely reject the hypothesis that the expert is statistically accurate. If, instead, four realizations fell outside the fifth to ninety-fifth percentile intervals, then the P-value drops to 0.01, and if five realizations fall outside the intervals, the P-value drops again, to 0.002.

[7] The percentiles elicited are also called quantiles.

Measuring Statistical Accuracy

The P-value is a measure for assessing the goodness of fit between a statistical hypothesis—in this case, the expert's assessments—and the data. The P-value is the cornerstone of validation in the classical model, and it is referred to as the expert's statistical accuracy score.[8] Statistical accuracy scores range from 0 to 1, with higher scores being better. The P-value is not used to accept or reject experts. Rather, the P-value is used because it is a familiar measure of the degree to which a statistical hypothesis is supported by data.

Measuring Information

In the classical model, statistical accuracy is one characteristic of an expert's performance. The second characteristic is how much information is provided in the expert's assessments. For example, consider an uncertain quantity that can vary between zero and one (e.g., the proportion of a population that has been exposed to a hazard). A ninety percent credible range from 0.03 to 0.90 is less informative than a credible range from 0.05 to 0.10. Among statistically accurate assessments, narrower informative assessments are more useful than wide, uninformative assessments. Thus the classical model assigns each expert an information score, which is based on the density of the expert's assessments relative to a background distribution.[9] To illustrate, if Expert A provides a narrower interval for her ninety percent credible range than Expert B, then Expert A's assessment is more densely concentrated and her information score is higher. The information score is scale invariant (i.e., the score does not depend on the units of the items) and insensitive to a distribution's tails. For more detail on the calculation of both the information and statistical accuracy scores, see, e.g., Cooke (1991), Cooke and Goossens (2008), Koch et al. (2015), and Wittmann et al. (2015).

Combining Experts' Judgments

The statistical accuracy and information scores are multiplied to create an expert's combined score. Information scores typically vary between experts by a factor of about three in a given study, whereas statistical accuracy scores vary over several orders of magnitude (Cooke and Goossens 2008; Colson and Cooke 2017). As a result, the combined score is influenced more by an expert's statistical accuracy than by her information. This means that in the overall combined score an expert cannot overcome poor statistical performance by providing very narrow credible ranges (and thus having a high information score).

Combined scores serve as the mechanism for producing performance-based weights for combining the experts' assessments (Cooke 1991; Cooke and Goossens 2008). In general, expert weights should be determined according to a *strictly proper scoring rule*, meaning that

[8] Statistical accuracy is sometimes also referred to as "calibration" (e.g., Cooke and Goossens 2008; Quigley et al. 2018). However, this can lead to the mistaken belief that analysts "calibrate" or adjust experts' assessments in the same way that scientists calibrate their instruments.

[9] The background distribution is chosen by the analyst, with the standard choice being a uniform (or log-uniform) distribution that includes the range of values provided by the experts, the actual value for that question (if an actual value exists), and an overshoot that extends the distribution's range on either side by plus and minus ten percent. For each expert and item, a density is fit to the background distribution, which adds the least information to the background while complying with the experts' quantile assessments.

an expert will maximize her expected score if and only if she states her true beliefs (Cooke 1991). The classical model is based on an *asymptotic* strictly proper scoring rule,[10] whereby an optimal cutoff value for statistical accuracy is used.[11] Experts with statistical accuracy scores below the cutoff value receive a weight of zero, which means their assessments are unweighted in the subsequent combination of the experts' judgments.

A combination of expert assessments is called a *decision maker*. The classical model can produce two types of performance-based decision makers: the global weight decision maker and the item weight decision maker. The global weight decision maker assigns each expert a constant weight for all items, based on the experts' average information score over all calibration questions. The item weight decision maker, however, assigns each expert a weight that varies for each item, based on the experts' information scores for that given item. With item weights, if an expert has statistical accuracy above the cutoff threshold and generally has high information scores but has a low information score on one question, she will receive less weight on that question relative to the others.

The performance of these two combined assessments can also be validated based on the calibration questions. In practice, in the classical model, analysts typically compare one or both performance-weight (PW) decision makers to a combination that weights all the experts equally, regardless of their performance. This is known as the equal-weight (EW) decision maker.

Calibration Questions

In practice, the classical model validates experts' assessments and scores expert performance through calibration questions, which are also called seed questions. Calibration questions serve three purposes in the classical model: they validate expert performance, enable performance-based expert weighting, and provide a mechanism for evaluating different combinations of the experts' assessments (Cooke et al. 2014). Although experts are not expected to know the precise true values for calibration questions, they are expected to provide reasoned quantifications of their uncertainty.

Calibration questions should not be general knowledge or almanac-type questions (e.g., What is the population of Venice, Italy?), as experts do not perform better on these questions than nonexperts and an expert's performance on these types of questions does not predict her performance on the variables of interest questions in her field (Cooke, Mendel, and Thijs 1988). Source data for calibration questions can include forthcoming reports, unpublished data sources, or analysis of unique combinations or subgroups of existing data.[12]

Finding potential calibration questions is a challenge that requires a strong understanding of the elicitation's subject matter. Calibration questions should be closely related to the variables of interest and should reflect the field(s) of the participating experts. Like the variables of interest, these questions need to be well specified to ensure that the experts

[10] This means that for large sets of assessments, an expert will maximize her expected score by expressing her true beliefs.

[11] For full mathematical details, see Cooke (1991) and the online supplementary materials for Colson and Cooke (2017).

[12] Quigley et al. (2018) discuss strategies for identifying calibration questions and include examples from several elicitation protocols.

interpret and understand the questions similarly. Because the calibration questions enable the performance-based expert weights, they must be high quality for the study to be credible. If reviewers or decision makers in the field do not believe that performance on the calibration questions is an appropriate indicator of ability to assess the variables of interest, then they will reject expert weights that are based on those questions.

Alternative Approaches to Aggregating Expert Judgments

The classical model combines experts' assessments in a linear opinion pool with weights determined by expert performance, thus incorporating an element of validation. Other types of weights and combination mechanisms have also been considered, some of which are reviewed here.

Likelihood-Based and Social Network Weights

Cooke, ElSaadany, and Huang (2008) compared classical model–based performance weights with two alternative approaches: likelihood-based weights and social network weights. They state that likelihood weights should be avoided because they rely on an improper scoring rule: they reward an expert for overstating her confidence rather than for expressing her true beliefs. Moreover, they find that the likelihood weight–based combinations do not outperform the classical model's weights in terms of statistical accuracy, information, or even likelihood scores. They also find that social network weights, which are based on an expert's scientific citations, can result in combined assessments with "unacceptably low" statistical accuracy. Focusing on citations may also restrict the expert pool to include only those experts with an established track record of publications, and the authors did not find support for such a restriction.

Peer Weights

Burgman et al. (2011) and Aspinall and Cooke (2013) explored the use of peer weights rather than performance weights, an idea that dates back to DeGroot (1974). Peer weights sound compelling: experts in a field are familiar with each other and each other's work. Thus, if experts agreed on whose opinion should receive the most weight, calibration questions would not be needed. However, the data from the few studies that examine this approach warn against it. For example, Aspinall and Cooke (2013) found that peer rankings, in which the experts rate each other's expertise, do not correspond well to rankings according to performance. In a review of six expert panels, Burgman et al. (2011) found that peer rankings correlate with the traditional hallmarks of expertise—for example, years of experience and publication record—but not with actual expert performance. This highlights the need for a method that emphasizes testing and validation, like the classical model.

Averaging Experts' Quantiles versus Averaging Distributions

Beyond the use of alternative linear pooling mechanisms, one additional issue that has received attention in the recent expert elicitation literature is whether analysts should average experts' quantiles (i.e., the fifth, fiftieth, and ninety-fifth percentiles elicited from the experts) or the experts' distributions. Although most approaches—including the classical model—average distributions, Lichtendahl, Grushka-Cockayne, and Winkler (2013) argue that averaging quantiles may be the preferred approach because it concentrates more of the resulting distribution around the median, thus reducing the alleged problem of wide uncertainty distributions in combined assessments.[13] Although this approach may sound appealing, we show later that averaging quantiles rather than distributions has substantial performance costs.

Median versus Mean of Quantiles

Hora et al. (2013) propose an aggregation of expert assessments that takes the median rather than the mean of the elicited quantiles. They show that this method performs well when experts are well calibrated and independent, but that mean-based approaches do better when those assumptions are relaxed.[14]

A Case Study of the Classical Model

To illustrate how the classical model has been applied and its policy relevance, we examine an application related to managing invasive species in the U.S. Great Lakes.[15] Researchers used the classical model to investigate the impact of existing invasive species on the economic value of ecosystem services (Rothlisberger et al. 2012), the possible future ecological impact of an Asian Carp invasion in the Great Lakes (Wittmann et al. 2015), and the effectiveness of various strategies to prevent the establishment of Asian carp in the Great Lakes (Wittmann et al. 2014). In addition, Zhang et al. (2016) use outputs from Wittmann et al. (2015), along with other data, to further consider the impact of Asian carp in the Great Lakes. Lodge et al. (2016) discuss other methods that have been used to understand the risks of invasive species, emphasizing the potential role of expert elicitation in supplementing gaps in empirical data and existing models. We focus here on Rothlisberger et al. (2012) to illustrate how the classical model can be directly applied to environmental policy decision making. More specifically, we describe the study's purpose and design, the elicitation process, how the experts performed, and the study's results.

Study Background and Setting

Rothlisberger et al. (2012) estimated the economic costs associated with the second wave of nonindigenous invasive species into the Great Lakes following the opening of the St. Lawrence Seaway.[16] The economic benefits of the seaway are easily quantified as

[13] Averaging quantiles is the same as harmonically weighting the experts' densities (Bamber, Aspinall, and Cooke 2016; Colson and Cooke 2017).

[14] Later we will examine whether the assumption of well-calibrated experts holds in data from thirty-three structured expert judgment studies conducted between 2006 and March 2015.

[15] This work was led by researchers at the University of Notre Dame and funded by the U.S. National Oceanic and Atmospheric Administration (NOAA) and the U.S. Environmental Protection Agency (EPA).

[16] The St. Lawrence Seaway is a system of locks, canals, and channels that permit oceangoing vessels to travel from the Atlantic Ocean to the Great Lakes as far inland as the western end of Lake Superior.

transportation cost savings. However, the economic costs of degraded ecosystem services are not directly observable and thus may be discounted or ignored in policy discussions. Expert elicitation allowed the researchers to conduct a policy thought experiment: what would be the value of ecosystem services in the U.S. Great Lakes region if the St. Lawrence Seaway had never opened and hence there were no shipborne invasive species? This enabled them to calculate the costs of invasive species without requiring long-term data on the value of ecosystem services and all of the potential confounding variables before and after the St. Lawrence Seaway was opened.

The Elicitation

Rothlisberger et al. (2012) considered damages related to four ecosystem services: commercial fishing, sportfishing, wildlife viewing, and raw water usage. The experts were asked to quantify their uncertainty about these services in the current (i.e., invaded) condition and in a hypothetical counterfactual (i.e., if shipborne invasive species were not present but all other factors remained unchanged). The expert estimates were then converted into dollar values, thus providing an estimate of the economic loss from degraded ecosystem services.

Rothlisberger et al. (2012) identified experts through a review of the relevant scientific literature and recommendations from other senior researchers in the field. Experts included academics, consultants, and government scientists who worked on issues related to ecosystem services in the Great Lakes. Nine experts participated, and they assessed thirteen calibration questions and forty-one variables of interest. Calibration questions asked about the following ecosystem services in 2006: pounds of commercially landed fish, angler days of sportfishing, expenditure on sportfishing, participant days of wildlife watching, and costs to raw water users. True values of these questions were unknown at the time of the elicitations, but the values were later released in annual reports from the U.S. Fish and Wildlife Service and the U.S. Geological Survey. The variables of interest concerned predictions of the same items in 2025 and estimates of what the values would have been in 2006 and 2025 if shipborne invasive species had never been introduced into the Great Lakes (i.e., if the St. Lawrence Seaway had never opened).

Expert and Decision-Maker Performance

Rothlisberger et al. (2012) found that the statistical accuracy of the experts varied over several orders of magnitude, from 0.45 to 1.2E-9.[17] This highlights the importance of *external* validation of the experts: although all nine experts had extensive subject matter expertise, as indicated by their positions and professional qualifications, they differed in their ability to make statistically accurate probabilistic statements quantifying their uncertainty. Assigning each expert equal weight, regardless of individual performance, produces a decision maker with a statistical accuracy score of 0.044. This means that the hypothesis that the EW decision-maker's probability statements are statistically accurate given the actual values of the thirteen calibration questions is below the traditional five percent level for rejecting statistical hypotheses. The PW decision maker had a much higher statistical accuracy score (0.928),

[17] The scores of the individual experts and the EW and PW combinations are presented in Appendix table 1.

indicating its better statistical performance in assessing the calibration questions. The PW decision maker assigns positive weight to only two experts (all of the other experts are unweighted), both of whom displayed good statistical accuracy.

Rothlisberger et al. (2012) found that even though seven of the experts are unweighted in the PW decision maker, the median estimated impacts from the PW decision maker were similar to those from the EW decision maker. However, the 90% credible ranges of the EW decision maker were, on average, 34% larger than those of the PW decision maker, indicating that the performance-weight decision maker was more informative.

Study Results

Based on the PW combination of the experts, Rothlisberger et al. (2012) estimated that the median annual loss to commercial fishing due to the impact of shipborne invasive species was $5.3 million. The experts estimated the impact on sportfishing to be greater, but also more uncertain; the estimated median annual loss was $106 million, but there was a five percent chance the impact exceeded $800 million. The median annual costs due to the impact of biofouling (i.e., the accumulation of organisms on underwater equipment or surfaces) on raw water usage, aggregated over all U.S. Great Lakes facilities, was estimated to be $27 million. Rothlisberger et al. (2012) dropped the costs associated with wildlife viewing from the analysis because the experts' uncertainty about the impact of invasive species on wildlife viewing was extremely wide. Thus the total ecosystem service losses were estimated to be $138.3 million annually.

Next Rothlisberger et al. (2012) estimated the cumulative economic loss associated with these annual impacts. Assuming a three percent discount rate and that the cost of invasive species over the next fifty years increases at the same rate as it has in the past, they extrapolate from the structured expert judgment and estimate that over the next fifty years, preventing future shipborne invasions would avoid more than $1.45 billion in cumulative losses in the United States from degraded ecosystem services. They also extrapolate from the estimated annual transportation cost savings associated with the St. Lawrence Seaway (see Taylor and Roach 2009) and find that the total cumulative transportation savings over fifty years is $1.41 billion. Thus the ecological services losses exceed the transportation savings. Rothlisberger et al. (2012) find that the most extreme countermeasure to shipborne invasive species in the Great Lakes—completely closing the St. Lawrence Seaway to oceangoing ships—would not produce net benefits for forty-nine years. Less extreme (and thus more realistic) options, such as deep ocean ballast water exchange, have shorter payback times.

In summary, Rothlisberger et al. (2012) present new information on the present and future costs of invasive species in the Great Lakes. Their estimates, which are possible only through the use of structured expert judgment, allow for a more robust discussion of the different policies available to control invasive species entering the Great Lakes through the St. Lawrence Seaway.

Applications and Performance of the Classical Model

We next review expert performance in recent applications of the classical model. First, we describe a dataset of thirty-three applications conducted between 2006 and March 2015.

Then we analyze the performance of the experts from these studies and compare the performance of three different weighting schemes: equal weighting, performance-based weighting (i.e., the weights produced by the classical model), and harmonic weighting.

Applications of the Classical Model

Cooke and Goossens (2008) compiled, analyzed, and publicly released the forty-five known applications of the classical model conducted through 2006. These data, which included studies with numbers of experts ranging from four to seventy-seven and numbers of calibration questions ranging from five to fifty-five, allowed other researchers to further study the classical model and consider how to validate and combine expert judgments (see e.g., Clemen 2008; Lin and Bier 2008; Lin and Cheng 2008; Lin and Huang 2012; Eggstaff, Mazzuchi, and Sarkani 2014). The pre-2006 data, however, include studies from the initial days of the classical model, before study design and procedures became more standardized. Moreover, use of the classical model has expanded greatly since 2006, driven in part by applications in high-visibility journals (e.g., Aspinall 2010; Bamber and Aspinall 2013), thus making the Cooke and Goossens (2008) dataset very out of date.

We are aware of thirty-three professionally contracted classical model studies that were performed between 2007 and March 2015.[18] All of these studies follow the general process outlined in a procedures guide for the classical model (Cooke and Goossens 1999) but are typically better documented than earlier studies. We conducted an analysis of these thirty-three applications of the classical model using Excalibur (Cooke and Solomatine 1992).[19]

In these thirty-three studies, 322 experts assessed between seven and seventeen calibration questions.[20] Studies included between four and twenty-one experts, and most studies had ten calibration questions (figure 1). More than one-third of the experts assessed ten calibration questions.

Performance of Experts

Although an expert's statistical accuracy score depends on the number of calibration questions assessed, for most experts the number of questions is similar, so we can make a rough comparison. We find that only eighty-seven of the experts have statistical accuracy greater than 0.05, which means we can reject the hypothesis that the other 233 experts provided statistically accurate assessments. More than half of the experts have statistical accuracy that is less than 0.005, and about one-third have scores of less than 0.0001 (figure 2). Again, this highlights the need for validation of expert judgments. Although all of the participants in these studies are established experts in their respective fields, their knowledge does not necessarily translate into statistically accurate probabilistic statements about unknown quantities.

[18] See the online supplementary materials for a list of the thirty-three applications of the Classical Model conducted between 2006 and March 2015, with references for the study documentation, where available.

[19] This is a freely available program now maintained by LightTwist Software and available at http://www.lighttwist.net/wp/excalibur.

[20] We excluded two experts who did not provide assessments for all the calibration questions in their respective panels. We also excluded one expert from the Koch et al. (2015) study because the expert was not included in the study's analysis.

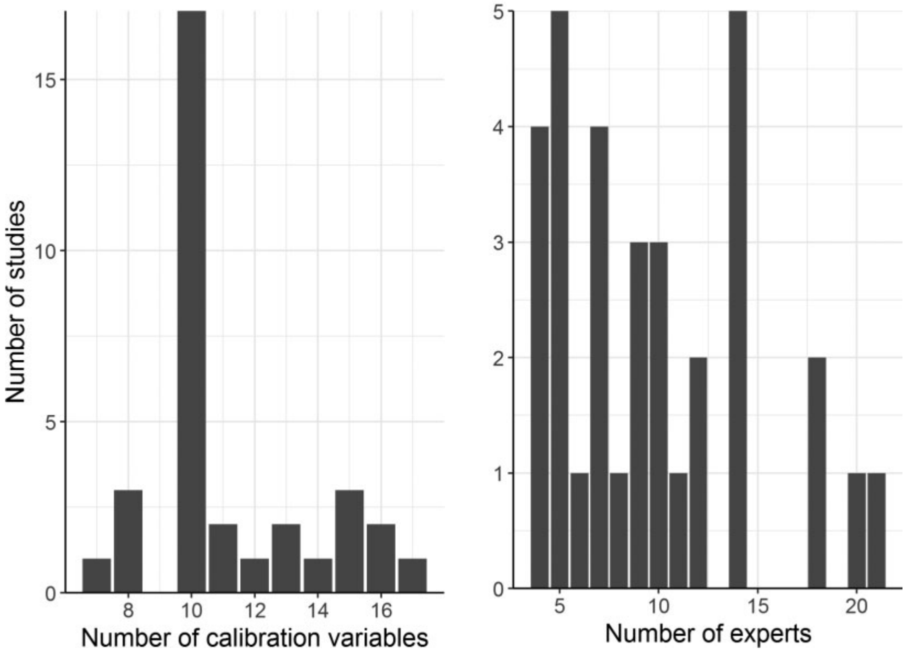

Figure 1 The number of experts and calibration questions in the thirty-three studies.
Source: Authors' calculations based on the 2006–March 2015 study dataset.

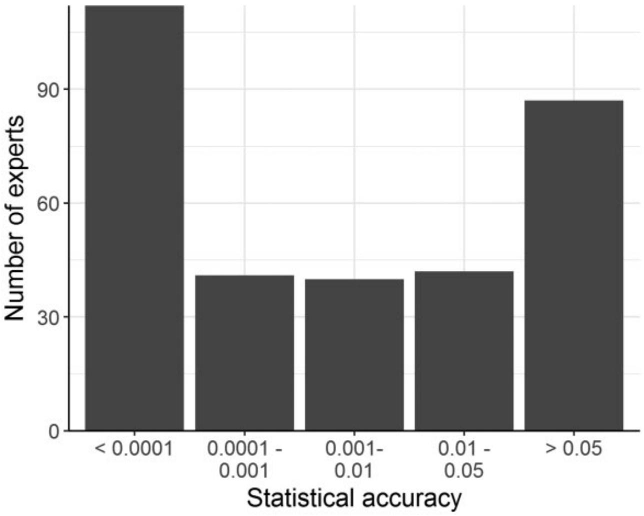

Figure 2 Distribution of expert statistical accuracy scores.
Source: Authors' calculations based on the 2006–March 2015 study dataset.

Although the majority of experts had statistical accuracy scores of less than 0.05, most studies had at least one statistically accurate expert—that is, an expert with a score greater than 0.05 (see figure 3). About twenty percent of the studies had no statistically accurate experts, and one study had twelve accurate experts. Approximately two-thirds of the studies had two or more statistically accurate experts. These results suggest that identifying the top

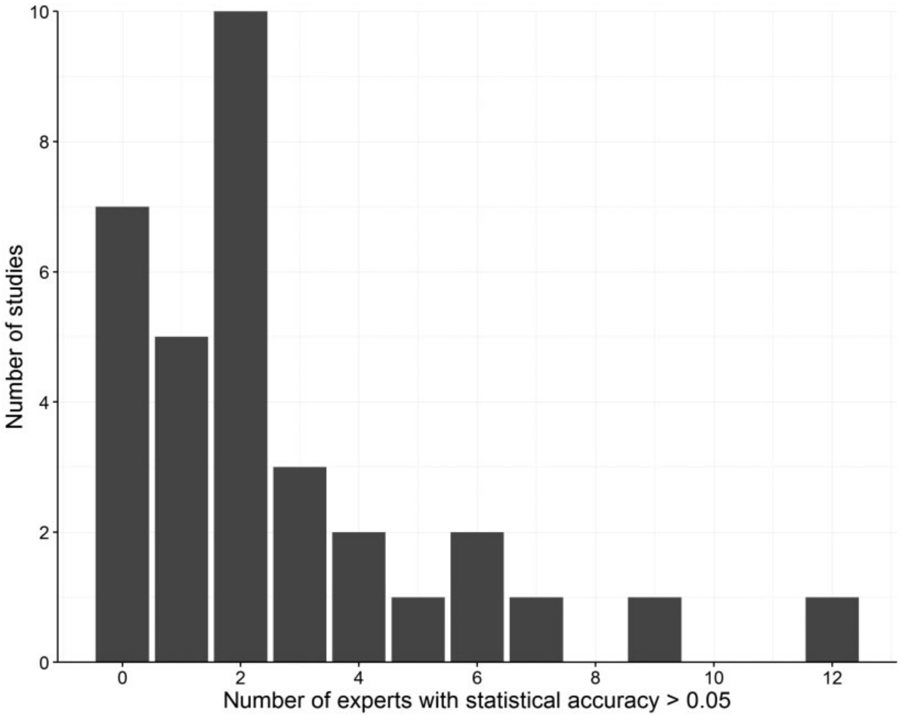

Figure 3 Number of statistically accurate experts per study.
Source: Authors' calculations based on the 2006–March 2015 study dataset.

experts in each study and relying on their judgments would be better than relying on expert judgments that have not been validated.

Performance of the Classical Model

As discussed earlier, the classical model identifies the optimal combination of judgments based on their performance (Cooke 1991; Cooke and Goossens 2008). In this subsection we will show that the scoring mechanism of the classical model further improves the quality of the expert data. More specifically, we compare the PW decision maker to the EW decision maker, which assigns equal weight to all of the experts without considering expert performance.

The PW decision maker can be based on item weights or global weights (as described earlier). In either case, the weight for each expert can range from zero (meaning the expert is unweighted in the PW decision maker) to one (meaning the expert is the only weighted expert in the PW decision maker; all other experts in that study are unweighted). The item weights decision maker gives nonzero weight to only one expert in thirteen of the thirty-three studies, while the global weights decision maker gives nonzero weight to one expert in fifteen cases. This means that only the expert with the highest statistical accuracy score in those studies is included in the PW decision maker; the other experts are all unweighted. This can mean that the classical model determined that only one of the experts in the study performed well or, more commonly, it can indicate that adding another expert's assessments to those of the best expert does not improve the overall performance; it just dilutes the assessments of the best expert.

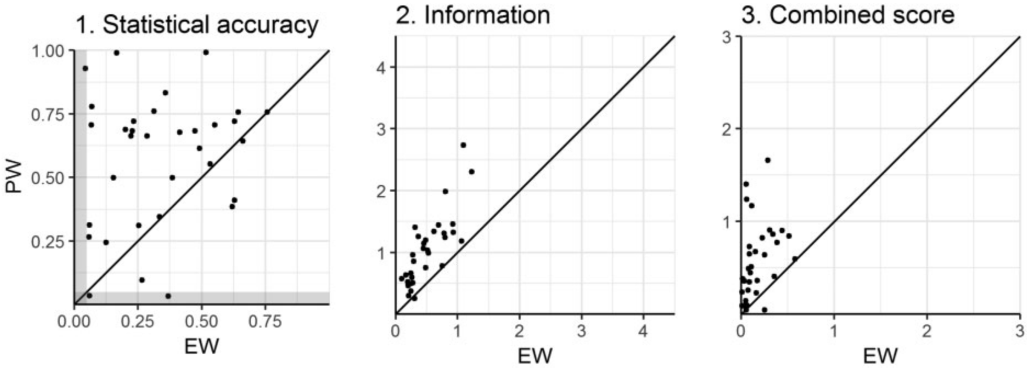

Figure 4 Statistical accuracy, information, and combined scores of the equal-weight (EW) and performance-weight (PW) decision maker for each of the thirty-three studies.
Notes: The shaded regions show statistical accuracy of less than 0.05. Each point is a single study and the $x = y$ line is drawn to show whether the score is higher for the EW or PW decision maker in that study.
Source: Authors' calculations based on the 2006–March 2015 study dataset.

The combined score of the PW decision maker based on global weights is greater than the combined score of the PW decision maker based on item weights in only six of the thirty-three studies. Because item weights are more commonly used in practice and typically are the same as or outperform global weights, unless otherwise specified, in the remainder of this section the PW decision maker will refer to the item weights decision maker.

Figure 4 compares the statistical accuracy, information, and combined scores of the PW and EW decision makers for each of the thirty-three studies. In twenty-six of the studies the statistical accuracy is greater for the PW decision maker than for the EW decision maker. This pattern is even more pronounced for information and the combined score; the PW decision maker outperforms the EW decision maker on information in thirty-two studies and on the combined score in thirty-one studies. These results indicate that performance-based weighting enables us to calculate a combination of the experts' judgments that is generally at least as statistically accurate as the equally weighted combination but is much more informative.

Performance of Harmonic Weighting

The classical model averages probability densities. Lichtendahl et al. (2013) argue for a different approach, which instead averages the experts' quantiles. This is equivalent to harmonic weighting (Bamber, Aspinall, and Cooke 2016). Implementing such harmonic weights without performance-based scoring does not require calibration questions, which means the elicitation process would be quicker and cheaper than in a full classical model study. Thus if harmonic weighting improved information relative to equal weighting while maintaining statistical accuracy, it would be an alluring alternative to performance-based weighting.

However, figure 5, which compares the performance of the harmonic-weight (HW) decision maker with the EW and PW decision makers,[21] cautions against such an

[21] Note that this HW decision maker, like the EW decision maker, is based on all of the experts' assessments, with no element of performance-based weighting.

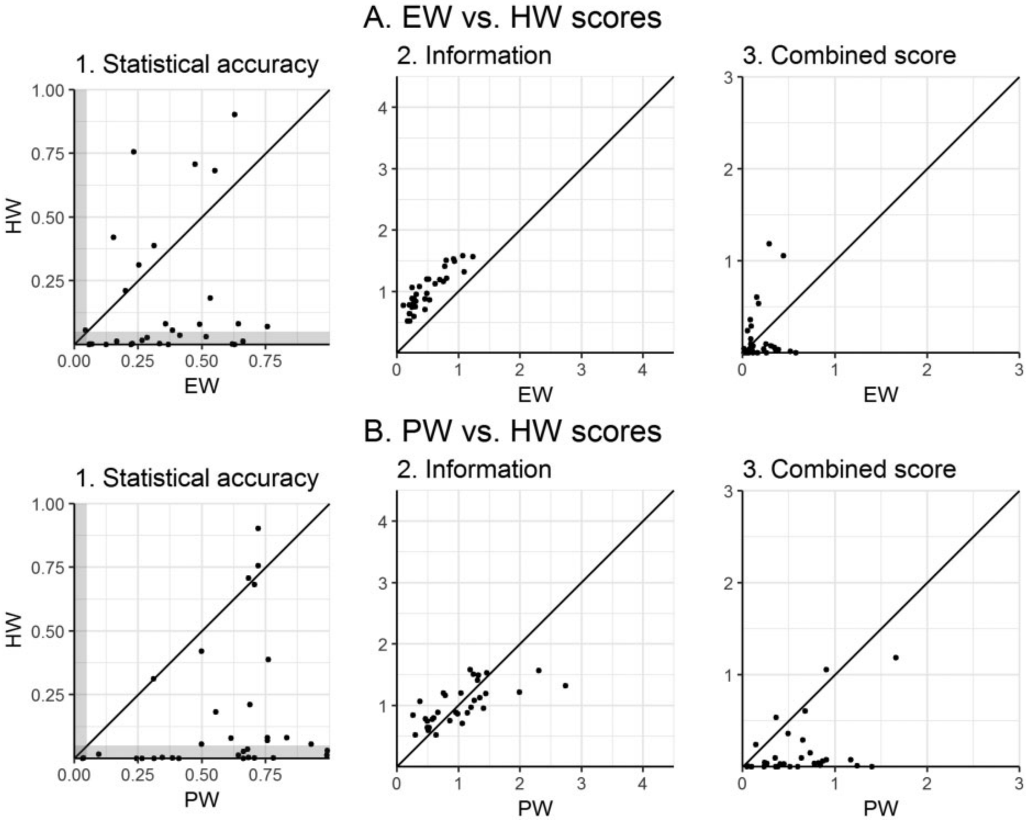

Figure 5 Statistical accuracy, information, and combined scores of the equal-weight (EW), performance-weight (PW), and harmonic-weight (HW) decision makers for each of the thirty-three studies.
Notes: The shaded regions show statistical accuracy of less than 0.05. Each point is a single study and the $x = y$ line is drawn to show whether the score is higher for the EW or PW decision maker in that study.
Source: Authors' calculations based on the 2006–March 2015 study dataset.

approach. Although the HW decision maker is more informative than the EW decision maker in all studies except one (figure 5, panel A2), its drop in statistical accuracy is drastic. In fact, the HW decision maker's statistical accuracy score is less than 0.05 in eighteen cases (Bamber, Aspinall, and Cooke 2016), less than 0.005 in twelve cases, and less than 0.0001 in four cases (figure 5, panel A1). In contrast, the lowest observed statistical accuracy for an EW decision maker is 0.04. Because the magnitude of the difference in statistical accuracy scores is larger than the difference in information scores, the EW decision maker has a higher combined score than the HW decision maker in all but ten studies (figure 5, panel A3). The PW decision maker compares favorably with the HW decision maker, with its combined score being higher than the HW decision maker's combined score in all but three studies (figure 5, panel B3). Thus, although harmonic weighting does improve information compared to equal weighting, it exacts a very high price in terms of statistical accuracy. In sum, if the goal of a weighting scheme is to yield good performance, it is important to measure and verify performance, as is done in the classical model.

Out-of-Sample Validation

The fact that the PW decision maker regularly has higher statistical accuracy, information, and combined scores than individual experts or the EW decision maker indicates the classical model's in-sample validity. This means that the classical model's weighting system performs well when it is evaluated against the calibration questions (i.e., the same data used to determine the weights). However, it is important to also examine how the model performs when it is tested on out-of-sample data. This section reviews recent work in this area.

Unfortunately, true out-of-sample validation is rarely possible for applications of the classical model because expert judgment is needed only when there are no data for the variables of interest. In fact, observations for the variables of interest were collected after the elicitation for only two of the forty-five cases in the Cooke and Goossens (2008) database, and none of the studies included in the data presented here have observations for the variables of interest. Thus, in lieu of evaluating true out-of-sample validity, the research has focused on techniques for *cross validation*, whereby a subset of the calibration questions is used to compute expert weights for a PW decision maker and the PW decision maker's performance is evaluated against the remainder of the calibration questions (i.e., the complementary subset). The first subset is called the *training set*, and the second subset is call the *test set*.

A common approach to cross validation is to remove one calibration question at a time, recalculate the combined scores and PW decision makers, and evaluate the decision makers' performance for the removed item (Clemen 2008; Lin and Cheng 2008, 2009).[22] In addition to the remove-one-at-a-time technique, researchers have investigated fifty–fifty splits of the calibration questions (Cooke 2008) and seventy–thirty splits (Flandoli et al. 2011). Eggstaff, Mazzuchi, and Sarkani (2014) take the most comprehensive approach, splitting a study's calibration question into all possible combinations of two subsets. This idea has also been applied in recent applications, which look at the data from an individual study (Cooke et al. 2014; Koch et al. 2015), but knowing how best to categorize a study's out-of-sample validity based on this procedure is difficult. Small training sets make it difficult to resolve differences in the statistical accuracy of the experts, and small test sets make it difficult to resolve differences between the EW and PW decision makers (Cooke 2014; Cooke et al. 2014). Colson and Cooke (2017) propose a summary measure that balances these issues and find that the PW decision maker outperforms the EW decision maker out of sample in twenty-six of the thirty-three 2006 to March 2015 studies. They also find that out-of-sample performance is correlated with the best and second-best experts' statistical accuracy scores but not with study characteristics that are easily within an analyst's control.[23]

[22] However, Cooke (2008, 2012a) noted that this procedure compounds the errors of increasing the weight of experts who assess the removed items badly, which results in significant bias.

[23] Colson and Cooke (2017) find that out-of-sample performance is not correlated with study characteristics such as the number of experts or calibration questions, whether the elicitations were done one on one or in plenary sessions, or whether three or five quantiles were elicited (typically the five, fifty, and ninety-five percentiles or the five, twenty-five, fifty, seventy-five, and ninety-five percentiles).

This recent work on the classical model's out-of-sample performance further demonstrates the validity of performance-based weighting of experts. Future work related to out-of-sample validation could be used to improve our understanding of the robustness of the results from a single study or to improve elicitation design.

Conclusions

Expert judgment plays an important (and unavoidable) role in risk management, uncertainty analysis, and decision making. Fortunately, techniques exist for collecting and validating expert judgments in a structured and scientifically sound manner. The classical model is one approach for validating and combining expert judgments, and it has been applied in more than one hundred expert panels to date, including the thirty-three single-panel applications conducted between 2006 and 2015 that we have discussed here. We have shown that the classical model's scoring system can identify the experts that are best able to quantify their uncertainty with meaningful probabilistic statements. Weighting the experts according to performance produces decision makers that consistently outperform combinations based on equal and harmonic weights in the thirty-three studies. Thus, for decisions or policies with a large potential impact on society, we would argue that the classical model and its validated assessments are the best tool for incorporating expert judgments when they are needed.

The classical model has been applied in a number of disciplines, but, as we have discussed, it is not appropriate in all circumstances. Even when expert judgment is needed, aggregating individual expert assessments into a single distribution may not be best in every application (Morgan 2014). More generally, expert judgment should not provide the final word on any issue; rather, it should guide future data collection, modeling, and analysis related to the topic.

Almost all of the thirty-three structured expert judgment studies were conducted in-person, either in one-on-one interviews or in plenary sessions. Recent World Health Organization elicitations, which we have not included here because they consist of a large number of overlapping expert panels, were conducted via *remote* elicitation (Aspinall et al. 2016). The experts in those panels did not perform as well as the experts in the panels reported here. This could be due to the remote elicitation. Thus, further work is needed to develop robust remote elicitation tools and to ensure that expert performance does not systematically decline with remote elicitation. Remote tools will make it easier to elicit, validate, and combine expert knowledge from around the world, potentially on an even larger scale than is currently done with the classical model.

Appendix Table 1 Scores and weights of the experts and combined assessments from the expert elicitation in Rothlisberger et al. (2012)

Expert or combination	Statistical accuracy	Information (variables of interest)	Combined score	Weight
1	4.03E-05	0.801	3.79E-05	0
2	0.0965	1.01	0.0677	0.3524
3	0.000117	1.52	0.000148	0
4	0.000117	1.51	0.000118	0
5	0.000747	0.578	0.000844	0
6	0.454	0.421	0.124	0.6476
7	0.000117	1.17	0.000116	0
8	4.86E-06	1.37	6.64E-06	0
9	1.91E-09	2.34	5.47E-09	0
Equal weight	0.0441	0.276	0.0135	–
Performance weight	0.928	0.424	0.240	–

References

Aspinall, W. P. 2010. A route to more tractable expert advice. *Nature* 463(7279):294–95.

Aspinall, W. P., and R. M. Cooke. 2013. Quantifying scientific uncertainty from expert judgement elicitation. In *Risk and Uncertainty Assessment for Natural Hazards*, ed. J. Rougier, S. Sparks, and L. Hill, 64–99. Cambridge: Cambridge University Press.

Aspinall, W. P., R. M. Cooke, A. H. Havelaar, S. Hoffmann, and T. Hald. 2016. Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. *PLoS One* 11(3):e0149817.

Bamber, J. L., and W. P. Aspinall. 2013. An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change* 3(4):424–27.

Bamber, J. L., W. P. Aspinall, and R. M. Cooke. 2016. A commentary on 'How to interpret expert judgment assessments of twenty-first century sea-level rise' by Hylke de Vries and Roderik SW van de Wal. *Climatic Change* 137(3–4):321–28.

Burgman, M. A., M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy. 2011. Expert status and performance. *PLoS One* 6(7):e22998.

Clemen, R. T. 2008. Comment on Cooke's classical method. *Reliability Engineering & System Safety* 93(5):760–65.

Clemen, R. T., and R. L. Winkler. 1999. Combining probability distributions from experts in risk analysis. *Risk Analysis* 19(2):187–203.

———. 2007. Aggregating probability distributions. In *Advances in Decision Analysis: From Foundations to Applications*, ed. W. Edwards, R. F. Miles, and D. Von Winterfeldt, 154–76. Cambridge: Cambridge University Press.

Colson, A. R., and R. M. Cooke. 2017. Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety* 163(July):109–20.

Cooke, R. M. 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science.* New York: Oxford University Press.

———. 2008. Discussion: Response to Discussants. In *Expert Judgement.* Special issue, *Reliability Engineering & System Safety* 93(5):775–77.

———. 2012. Uncertainty analysis comes to integrated assessment models for climate change. . .and conversely. *Climatic Change* 117(3):467–79.

———. 2012a. Pitfalls of ROAT Cross-Validation: Comment on Effects of Overconfidence and Dependence on Aggregated Probability Judgments. *Journal of Modelling in Management* 7(1):20–22.

———. 2014. Validating expert judgment with the classical model. In *Experts and Consensus in Social Science: Critical Perspectives from Economics*, ed. C. Martini and M. Boumans, 191–212. Cham, Switzerland: Springer International.

———. 2015. Messaging climate change uncertainty. *Nature Climate Change* 5(1):8–10.

Cooke, R. M., S. ElSaadany, and X. Huang. 2008. On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering & System Safety* 93(5):745–56.

Cooke, R. M., and L. L. H. J. Goossens. 1999. Procedures guide for structured expert judgment. EUR 18820. Delft, The Netherlands: Delft University of Technology. https://cordis.europa.eu/pub/fp5-euratom/docs/eur18820_en.pdf.

———. 2008. TU Delft expert judgment data base." *Reliability Engineering & System Safety* 93(5):657–74.

Cooke, R. M., M. Mendel, and W. Thijs. 1988. Calibration and information in expert resolution; a classical approach. *Automatica* 24(1):87–93.

Cooke, R. M., and D. Solomatine. 1992. *EXCALIBUR—Integrated System for Processing Expert Judgments, User's Manual Version 3.0.* Delft, The Netherlands: Delft University of Technology and SoLogic Delft.

Cooke, R. M., M. E. Wittmann, D. M. Lodge, J. D Rothlisberger, E. S. Rutherford, H. Zhang, and D. M. Mason. 2014. Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. *Integrated Environmental Assessment and Management* 10(4):522–28.

DeGroot, M. H. 1974. Reaching a consensus. *Journal of the American Statistical Association* 69(345):118–21.

Eggstaff, J. W., T. A. Mazzuchi, and S. Sarkani. 2014. The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering & System Safety* 121(January):72–82.

Flandoli, F., E. Giorgi, W. P. Aspinall, and A. Neri. 2011. Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety* 96(10):1292–1310.

Hora, S. C. 2007. Eliciting probabilities from experts. In *Advances in Decision Analysis: From Foundations to Applications*, ed. W. Edwards, R. F. Miles, and D. Von Winterfeldt, 129–53. Cambridge: Cambridge University Press.

Hora, S. C., B. R. Fransen, N. Hawkins, and I. Susel. 2013. Median aggregation of distribution functions. *Decision Analysis* 10(4):279–91.

Koch, B. J., C. M. Febria, R. M. Cooke, J. D. Hosen, M. E. Baker, A. R. Colson, S. Filoso, K. Hayhoe, J. V. Loperfido, A. M. K. Stoner, and M. Palmer. 2015. Suburban watershed nitrogen retention: estimating the effectiveness of stormwater management structures." *Elementa: Science of the Anthropocene* 3(July):000063.

Lichtendahl, K. C., Y. Grushka-Cockayne, and R. L. Winkler. 2013. Is it better to average probabilities or quantiles? *Management Science* 59(7):1594–611.

Lin, S.-W., and V. M. Bier. 2008. A study of expert overconfidence. *Reliability Engineering & System Safety* 93(5):711–21.

Lin, S.-W., and C.-H. Cheng. 2008. Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities? In *IEEE International Conference on Industrial Engineering and Engineering Management, 2008*, 425–29. New York: IEEE.

———. 2009. The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management* 4(2):149–61.

Lin, S.-W., and S.-W. Huang. 2012. Effects of overconfidence and dependence on aggregated probability judgments. *Journal of Modelling in Management* 7(1):6–22.

Lodge, D. M., P. W. Simonin, S. W. Burgiel, R. P. Keller, J. M. Bossenbroek, C. L. Jerde, A. M. Kramer, E. S. Rutherford, M. A. Barnes, M. E. Wittmann, W. L. Chadderton, J. L. Apriesnig, D. Beletsky, R. M. Cooke, J. M. Drake, S. P. Egan, D. C. Finnoff, C. A. Gantz, E. K. Grey, M. H. Hoff, J.

G. Howeth, R. A. Jensen, E. R. Larson, N. E. Mandrak, D. M. Mason, F. A. Martinez, T. J. Newcomb, J. D. Rothlisberger, A. J. Tucker, T. W. Warziniack, and H. Zhang. 2016. Risk analysis and bioeconomics of invasive species to inform policy and management. *Annual Review of Environment and Resources* 41:453–88.

Mellers, B., E. Stone, T. Murray, A. Minster, N. Rohrbaugh, M. Bishop, E. Chen, J. Baker, Y. Hou, M. Horowitz, L. Ungar, and P. Tetlock. 2015. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science* 10(3):267–81.

Morgan, M. G. 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America* 111(20):7176–84.

Morgan, M. G., and M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* Cambridge: Cambridge University Press.

National Academies of Sciences, Engineering, and Medicine. 2017. *Valuing Climate Changes: Updating Estimation of the Social Cost of Carbon Dioxide.* Washington, DC. https://www.nap.edu/catalog/24651/valuing-climate-changes-updating-estimation-of-the-social-cost-of.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities.* West Sussex, UK: John Wiley & Sons.

Oppenheimer, M., C. M. Little, and R. M. Cooke. 2016. Expert judgement and uncertainty quantification for climate change. *Nature Climate Change* 6(5):445–51.

Puig, D., and O. Morales-Nápoles. 2017. The accountability imperative for quantifying the uncertainty of emission forecasts: evidence from Mexico. *Climate Policy* Forthcoming.

Quigley, J., A. R. Colson, W. P. Aspinall, and R. M. Cooke. 2018. Elicitation in the classical model. In *Elicitation: The Science and Art of Structuring Judgement*, ed. L. C. Dias, A. M. Morton, and J. Quigley, chap. 2. New York: Springer.

Rothlisberger, J. D., D. C. Finnoff, R. M. Cooke, and D. M. Lodge. 2012. Ship-borne nonindigenous species diminish Great Lakes ecosystem services. *Ecosystems* 15(3):1–15.

Rowe, G., and G. Wright. 1999. The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting* 15(4):353–75.

Taylor, J. C., and J. L. Roach. 2009. Ocean shipping in the Great Lakes: an analysis of industry transportation cost savings. *Transportation Journal* 48(1):53–67.

U.S. Nuclear Regulatory Commission. 1975. *Reactor Safety Study.* WASH-1400, NUREG-75/014. Washington, DC: U.S. Nuclear Regulatory Commission.

Verdolini, E., L. D. Anadón, E. Baker, V. Bosetti, and L. A. Reis. 2018. Future prospects for energy technologies: insights from expert elicitations. *Review of Environmental Economics and Policy* 19(2):133–153.

Wittmann, M. E., R. M. Cooke, J. D. Rothlisberger, and D. M. Lodge. 2014. Using structured expert judgment to assess invasive species prevention: Asian carp and the Mississippi–Great Lakes hydrologic connection. *Environmental Science & Technology* 48(4):2150–56.

Wittmann, M. E., R. M. Cooke, J. D. Rothlisberger, E. S. Rutherford, H. Zhang, D. M. Mason, and D. M. Lodge. 2015. Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. *Conservation Biology* 29(1):187–97.

Zhang, H., E. S. Rutherford, D. M. Mason, J. T. Breck, M. E. Wittmann, R. M. Cooke, D. M. Lodge, J. D. Rothlisberger, X. Zhu, and T. B. Johnson. 2016. Forecasting the impacts of silver and bighead carp on the Lake Erie food web. *Transactions of the American Fisheries Society* 145(1):136–62.

Abstract

The inclusion of expert judgments along with other forms of data in science, engineering, and decision making is inevitable. Expert elicitation refers to formal procedures for obtaining and combining expert judgments. Expert elicitation is required when existing data and models cannot provide needed information. This makes validating expert judgments a challenge because they are used when other data do not exist and thus measuring their accuracy is difficult. This article examines the classical model of structured expert judgment, which is an elicitation method that includes validation of the experts' assessments against empirical data. In the classical model, experts assess both the unknown target questions and a set of calibration questions, which are items from the experts' field that have observed true values. The classical model scores experts on their performance in assessing the calibration questions and then produces performance-weighted combinations of the experts. From 2006 through March 2015, the classical model has been used in thirty-three unique applications. Less than one-third of the individual experts in these studies were statistically accurate, highlighting the need for validation. Overall, the performance-based combination of experts produced in the classical model is more statistically accurate and more informative than an equal weighting of experts. (*JEL*: C18, C19, C89)