cebra
Centre of Excellence for
Biosecurity Risk Analysis

Comparability of Elicitation Approaches: Literature review & Experimental design

Final Report

Anca Hanea, CEBRA, The University of Melbourne

Andrew Robinson, CEBRA, The University of Melbourne

Executive Summary

Expert elicitation has been used by the Risk Return Resource Allocation (RRRA) team in the Department of Agriculture to obtain data on elements of the biosecurity system that are highly uncertain (i.e., where quantitative information is scarce or absent). Elicitation protocols have varied over time. This raises the question of whether results can be compared across different protocols. The intuitive answer is: no, and this intuition is informed by insights in the complexity of an expert elicitation. However, to what extent this intuition is correct remains to be investigated. Rather than contrasting all possible variations in elicitation protocols, we focus on one such variation, namely *the forum in which the experts interact*, which is remote or in face-to-face workshops. We consider structured protocols only, as the unstructured ones can be classed as unscientific and are impossible to contrast. The topic of the current investigation therefore to ask: are results obtained using a face-to-face structured elicitation comparable with the results obtained using a remote one? We further restrict our problem space to the biosecurity application area.

We first undertook a literature review of existing elicitation protocols used in biosecurity. Less than half of these papers were qualified as using structured protocols. None of the reviewed papers compared face-to-face elicitation results with remote elicitation. Moreover, many of the reviewed papers used hybrid (combined remote and face-to-face) protocols. This suggests the need of a three-way comparison, face-to-face, remote, and hybrid.

In the absence of evidence that supports the comparison of protocols, we propose an experiment for comparing performance of these variations as embedded in the protocol used in the Australian biosecurity research (namely the IDEA protocol). Comparing performance of approaches/treatments in an experiment requires validation data (questions estimated by experts for which the answers become available post-elicitation). In the analysis, the cost of the elicitations (including finding validation data), and the accuracy and calibration of judgements from various elicitations should be contrasted. However, given the complexity of such experiments, we recommend that the RRRA team source calibration questions, expert estimates and true answers from the previous elicitations and use these in a statistical (regression) model to try to compare the protocols or to inform a statistically valid experimental design.

1 INTRODUCTION

In order to fill cognitive gaps, experts can estimate quantities (parameters of models), decide the forms of cause–effect relationships, and predict the outcomes of management interventions, all of which are uncertain. Here, we restrict our attention to the elicitation of parameters and their associated uncertainties from experts. How these are best elicited and combined across experts is critical to a decision process, as differences in the efficacy and robustness of elicitation and aggregation methods can be substantial (e.g., Morgan, 2015; O'Hagan et al., 2006; Clemen and Winkler, 1999). These differences may arise from the different steps analysts choose to take along the way in terms of the definition of expertise, the number of experts used in an elicitation, the diversity of the expert group, the choice of questions and their format, the platform used for the elicitation (remote or face-to-face), the extent of feedback provided to experts, the way and extent to which experts interact prior and during the elicitation, etc. Many of these elements and their effect on the reliability of experts' judgements were investigated in isolation and for different protocols throughout many years. Theoretical studies of how to coherently use expert judgement go back to the 1960s and arguably a century or two before that (e.g., French, 1985; Cooke, 1991).

A comprehensive discussion on all of the elements mentioned above is out of scope, however emphasising the importance of a few of these elements and the way they shape elicitation protocols will follow. We will first discuss what and how we chose to ask for experts' judgements, and how should different judgements be combined after their elicitation from a diverse group of experts. We then enumerate and motivate the elements of a formal/structured elicitation protocol. This discussion will set the context and standards for what one should expect from an expert elicitation that can be deemed as scientific. Using this as a starting point, and assuming a one dimensional variation of a structured elicitation protocol, we are interested in the effects of such variation on the reliability of the elicited estimates.

1.1 Eliciting uncertainty

There are two ways in which uncertainty is acknowledged and modelled through expert elicited data. One is to ask a panel of experts about point estimates of (theoretically) measurable variables and

then take the variability between experts' answers as a measure of uncertainty. Another is to ask a panel of experts about their subjective distribution associated with a (theoretically) measurable variable and aggregating experts' uncertainties/distributions. The latter approach, even though more demanding, is recommended as it obviously captures much more information and a better description of uncertainty within and between experts.

Quantifying subjective distributions, or uncertainty through expert elicitation protocols may take two forms: eliciting probabilities, or eliciting values of a continuous variable corresponding to different percentiles. Eliciting probabilities can take several forms depending on the variables involved in the model. Those may correspond to:

- probabilities of event occurrences (e.g., Pr(pest entry), Pr(pest establishment));
- probabilities of various states of discrete variables, conditional probabilities for combinations of discrete states of variables (e.g., conditional probability tables (CPTs) in Bayesian networks). These discrete variables can be genuinely discrete, or discretized variables that are theoretically measured on a continuous scale;
- in the case of a continuous variable, the probability of a given value (i.e., values on the y-axis of a plot representing the probability density function (PDF) or the cumulative distribution function (CDF) of a variable.)

Some argue that answering questions about probabilities (during an elicitation) is more difficult than answering questions about quantities because a "probability doesn't exist". Hence, when continuous variables are modelled and their distribution needs to be elicited from experts, we recommend eliciting values of these continuous variables corresponding to a finite number of different percentiles — typically three — e.g., values on the x-axis of a plot representing the CDF of a variable.

What we appeal to, when asking about probabilities, are the different interpretations of probabilities available to the experts; such probabilities can be either thought of as a subjective representation of a rational preference, or as a relative frequency. An extra complication associated with eliciting probabilities is the quantification of the imprecision in such estimates within a probabilistic framework. To account for such imprecision, upper and lower bounds are requested, in addition to a

best estimate for a probability. When the probabilities can be interpreted as relative frequency, the bounds can be interpreted as percentiles of the expert subjective probability distribution. However, when the relative frequency interpretation is not appropriate (i.e., when the probability of a unique event is elicited) the bounds may be criticised for lacking operational definitions (in a "classical" probabilistic framework). However, many protocols argue that the main reason to elicit bounds in such cases is to improve thinking about the best estimates (O'Hagan et al., 2006; Hanea et al., 2018b).

We suggest to *always* ask for bounds. If the probabilities in question do represent relative frequencies, then we suggest formulating the questions in terms of relative frequencies, treating these as continuous variables and ask for percentiles of the experts' subjective distribution. In this way the richest possible elicitation is achieved, and that allows for more in-depth robustness and sensitivity probabilistic analyses. If the probabilities needed for the model correspond to one-off events, asking for bounds provokes counter-factual thinking and helps with the estimation of the best estimate. However, we then recommend using only the best estimate in further probabilistic analysis.

A summary of what can be elicited during an expert elicitation in order to quantify and model uncertainty is provided in the first two columns of Table 1. The third column of the table summarizes the choices one has in terms of eliciting more or less layers of uncertainty. The last column gives an indication of the various interpretations of probabilities that can and should be used in order to inform the formulation of the questions and the operationalization of concepts.

Table 1: Elicited estimates when quantifying uncertainty

Variable type	Estimate elicited	Treatment of uncertainty	Interpretation of probability
Discrete	—probability of event occurrences	point estimate / point estimate & bounds	subjective[1] / frequentist
	— (conditional) probabilities of discrete states	point estimate / point estimate & bounds	subjective / frequentist
Continuous	—probability of a given value	point estimate	subjective
	—percentiles of a distribution	credible interval	subjective distribution of the expert

The approach that modellers end up choosing (among eliciting point estimates, probabilities represented as frequencies, eliciting parameters of distributions, percentiles, or intervals) depends on the context, the needs of the chosen probabilistic model where the elicited estimates will be embedded in, the resources available, and the familiarity of the experts with probabilistic concepts. Hence some of the limitations relate to the *human nature* of the elicitation exercise, while others are *technical*. Often it is the interplay of the two that makes things even more complicated. Even though more than one option can be chosen, it is unusual for elicitations to take various forms for purely theoretical investigations. To our knowledge there are no experiments designed to compare and contrast these alternatives.

1.2 Aggregating multiple expert estimates

Whatever we chose to elicit, by agreeing that we are going to use a panel of experts, we must assume that the elicitation outcome will consist of a set of estimates, distributions, etc., that will need to be aggregated for later use.

There are two main ways in which experts' judgements are pooled (Clemen and Winkler, 1999): using *behavioural aggregation*, which involves striving for consensus via discussion (O'Hagan et al., 2006), or using *mathematical aggregation* which provides a more explicit, auditable and objective approach to aggregation. A weighted linear combination of opinions is one example of such aggregation. Equal weighting is often used mostly because of its simplicity (no justification for weights is required). Evidence also shows that the equal weighting scheme frequently performs quite well relative to more sophisticated aggregation methods (e.g., Clemen and Winkler, 1999), but not always (Cooke and Goossens, 2008; Cooke, 2015).

We recommend differential weighting based on prior expert performance on similar tasks. A few examples from the literature support this recommendation: for example the authors of Woudenberg (1991); Burgman et al. (2011); Cooke et al. (2008) used self-ratings, peer-ratings, and citation indices to formulate weights and found that the performance of an aggregated opinion using such

[1]The probability itself is a quantification of uncertainty. Second order uncertainty may be expressed through bounds. These bounds can be operationalized as percentiles of the experts' subjective distribution on an unknown relative frequency. If the probability is given for an one off event, then the bounds cannot be interpreted within the classical probability theory framework.

weights performs poorly in terms of accuracy and informativeness. Probably the most well known and widely used version of a differential weighting scheme is the Classical Model or Cooke's model (CM) for SEJ (Cooke, 1991), which uses calibration variables[2] to derive performance based weights proportional to how calibrated, accurate and informative the experts estimates are.

Mixed SEJ protocols combine behavioural and mathematical aggregation methods (Ferrell, 1994). The most common mixed protocol is the Delphi protocol (Rowe and Wright, 2001), in which experts receive feedback over successive questionnaire rounds, in the form of other group members' judgements. Experts remain anonymous and do not interact with one another directly. Instead, a facilitator provides feedback between rounds. As originally conceived, the Delphi protocol strives to reach consensus after a relatively small number of rounds (Dalkey, 1969), though in modern usages achieving consensus is not necessarily the primary aim (e.g., von der Gracht, 2012).

The IDEA protocol (Hanea et al., 2016; Hemming et al., 2018a), developed by ACERA/CEBRA (and used in the Australian biosecurity research), synthesizes specific elements from the aggregation approaches described above. IDEA is so-called because it encourages experts to Investigate, Discuss, and Estimate, and concludes with a mathematical Aggregation of judgements. It is a Delphi-like protocol in that experts give individual judgements over subsequent rounds, and facilitators provide feedback. In contrast to the traditional Delphi, IDEA does not seek consensus and can not always ensure full anonymity. A diverse group of experts first answers questions without engaging in discussion. Experts are then provided with the judgements of their peers and have the opportunity to endorse agreements and discuss differences of opinion (unlike Delphi), allowing people to reconcile the meanings of words and context (e.g., Carey and Burgman, 2008). Facilitators encourage and moderate discussion between rounds. All the steps in the protocol were informed by research in psychology, decision theory, applied mathematics and previous expert elicitation research. These steps try to circumvent difficulties arising from both human and technical reasons.

[2]Calibration variables are variables taken from the experts' domain for which the true values are known, or will become known, within the time frame of the study Aspinall (2010).

1.3 Structured/Formal expert elicitation protocols

If expert opinion is used as scientific data, then it should be subject to the same kind of methodological rules for quality assurance that are applied to other types of empirical data (Cooke and Goossens, 2000; Cooke, 1991).

Several different elicitation protocols developed over the last decades have been deployed successfully (i.e., results were recognized as reliable and the protocols were recognised as scientific) in fields as diverse as political science, infrastructure planning, environmental sciences and volcanology (e.g., Cooke and Goossens, 2008; Aspinall, 2010; Aspinall and Cooke, 2013; O'Hagan et al., 2006; Bolger et al., 2014; Singh et al., 2018). Most follow thoroughly documented methodological rules, but they differ in several aspects, including the way interaction between experts is handled and the way an aggregated opinion is obtained from individual experts.

Numerous structured expert judgement (SEJ) protocols for uncertainty quantification are available. There is no single, best SEJ protocol; each has strengths and weaknesses depending on the specificity of the problem, how dispersed geographically the experts are, how familiar are the experts with expressing uncertainty numerically, what resources are available, etc. (e.g., Bolger et al., 2014; O'Hagan et al., 2006). Hanea et al. (2018b) identify the following elements that make an SEJ accountable, transparent and repeatable:

1. Asks questions that have clear operational meanings, i.e., something that in principle would be measurable if time and resources would permit experiments;
2. Follows transparent methodological rules, i.e., is traceable, repeatable and available to review;
3. Anticipates and mitigates some of the most important psychological and motivational biases[3];
4. The process is thoroughly documented; and
5. Provides opportunities for empirical evaluation and validation[4].

We suggest that the first four items are absolutely essential for a SEJ protocol for describing uncertainty, and that it is imperative that the elicited information is expressed quantitatively.

[3]There are hundreds of heuristics that people use when reasoning under uncertainty. These lead to biases that can not be eliminated but can be mitigated. Most elicitation protocols try to mitigate against anchoring, availability, overconfidence and group think

[4]In practice this requires calibration questions, which ideally are question for which the true answers will soon become available.

Qualitative elicitations invite vagueness and trigger biases.

The empirical control requirement (formulated by the 5th item) is essential to at least one SEJ protocol (the CM), and the proponents of CM argue that it is essential to any elicitation protocol which calls itself *structured*. It is this requirement that justifies the use of calibration variables, providing an empirical basis for validating expert' judgements that is absent in other approaches. Using these the experts assessments can be checked for calibration, accuracy and informativeness, and so can be any combination we chose to use as a final aggregate answer. We note however, that other methods, lacking empirical control, but eliciting expert judgements in a structured manner, following a rigorous protocol, are also considered SEJ protocols (Bolger et al., 2014).

1.3.1 The IDEA protocol

The SEJ protocol that we recommend is the IDEA protocol. IDEA strives to comply with all the five items identified as essential to a SEJ. The IDEA protocol evolved to have a number of execution modes: (i) face-to-face, (ii) remote, and (iii) hybrid, i.e. first round is done remotely and the discussion and the second round happen during face-to-face workshops. These modes are described and detailed in Hemming et al. (2018a); Hanea et al. (2018a).

The most notable variation in implementing the IDEA protocol is its face-to-face versus remote implementation. This variation is shared by most of the elicitation protocols designed and used before Skype, Zoom, and other tools that can facilitate remote meetings were readily available. As mentioned earlier this variation occurred and was used in accordance with the needs of projects and the availability of resources. When time and other resources did not permit face-to face meetings, the remote version was used. No project has catered for the use of both implementations for purely academic comparisons. Because of this limitation, several pertinent questions remain unanswered: would all the benefits and carefully designed aspects of these protocols carry over in the same manner when the meetings are not held face-to-face? Will the engagement level stay the same? Will the responsibility and due diligence that experts feel change? Will there be aspects which will improve the quality of the elicitation when executed in remote mode?

These sort of questions motivated the present research. As mentioned in the beginning of Section 1, we are interested in how a one-dimensional variation of a structured elicitation protocol affects the reliability of the elicited estimates. The one dimensional variation is the remote versus face-to-face implementation of a SEJ and the particular application area we are focusing on is the biosecurity area.

We first undertook a literature review on existing elicitation protocols used in biosecurity. The hope was to inform the answers to some of the above formulated questions using existing research. However, none of the reviewed papers in the biosecurity domain compared face-to-face elicitation results with remote elicitation. We summarise our findings in Section 2, grouping references based on how they elicit uncertainty, on what aggregation method they employ, and what protocol they use (if specified), following the same lines of discussion as presented in Sections 1.1, 1.2 and 1.3.

In the absence of evidence that supports the comparison of protocols' variations we are interested in, in Section 4 we propose an experiment for comparing performance of these variations as embedded in the IDEA protocol. Section 5 gathers conclusions and recommendations.

2 Literature review

A search on Google Scholar (undertaken on 17/06/2019) using the terms "expert elicitation" and "biosecurity" in conjunction revealed 441 matches in the last 10 years (2010–2019). However, because the search looked through the entire text of the papers, many authors, applying expert judgement in other domains, but with "biosecurity" in their affiliation for example (e.g., CEBRA) were selected in the search. Also, many biosecurity applications that did not employ structured elicitation techniques were selected by this search.

One hundred papers were initially downloaded, and only 87 of those appeared relevant. None of the reviewed papers compared face-to-face elicitation results with remote elicitation, but the authors of Racicot et al. (2018) undertook a couple of elicitations, two years apart, on the same topic, using the same type of questions (about relative risks of different factors) and (some of) the same experts. The first elicitation was face-to-face and the second was done remotely. However, the authors did not comment on, or compared any aspects of the two elicitations.

Interestingly, many of the more recent references reviewed and qualified as using a structured protocols, elicited more than just point estimates (e.g., Singh et al., 2018; Peyre et al., 2016; Brookes et al., 2017; Hood et al., 2019; Hemming et al., 2018b; Muellner et al., 2018). The exceptions (where only point estimates are elicited) elicit scores, ranks, and percentages (Carmo et al., 2018), proportions, probabilities or ratio of probabilities (Gustafson, 2010), probabilities or values that are later fitted into mixed effects models (Barry and Lin, 2010), consequence severity on a discrete scale (Davidson et al., 2013), etc.

From the reviewed papers 27 out of 87 (31%) used the IDEA protocol or a similar Delphi-like approach. From these, 5 (18.5%) used a purely remote version, 12 (44.5%) used a face-to-face setting and 10 (37%) used a hybrid protocol. As mentioned before, the choice is often dictated by resources, the availability of on-line facilities and the number of experts that need to be involved.

2.1 Unstructured elicitations

Given that there is no strict definition of a structured protocol and we are only using the guidelines from Section 1.3, partitioning the references into *structured* and *unstructured* is challenging. Some references use an unstructured way of asking the questions, a poor reporting on whether uncertainty was elicited or not, but a structured protocol that ensures experts are consulted throughout the process (Jarrad et al., 2011a). Another example is Radia et al. (2013), where the protocol was structured but discrete choice was elicited together with self-elicited confidence scores which were then used as weights. Self-elicited confidence scores used as weights has no methodological justification.

Based on the recommendations formulated throughout Section 1, we will qualify as unstructured elicitations those which use qualitative scales to describe uncertainty and those which are not documented enough to allow scrutiny. A third criterion used to qualify protocols as unstructured (not discussed until now) is the mixture of SEJ with decision making modelling. Expert elicitation is sometimes badly integrated or confused with decision making where value judgements and preferences elicitation are required, rather than estimates of numbers and facts. Value judgements should not be elicited from experts. Structured preference modelling involves a distribution of preferences

across a group of stakeholders, rather than experts. The two should not be confused.

2.1.1 Qualitative

The authors of Delgado et al. (2016); Lohr et al. (2015); Van Klinken et al. (2016); Leger et al. (2017) use qualitative labels to describe uncertainty. Verbal labels that describe levels of uncertainty are understood differently in different context and by different people, which make comparisons between them difficult, and aggregation impossible to interpret. However, this did not stop the authors of Roche et al. (2015) from performing matrix calculations with qualitative matrix entries.

The authors of Brouwer et al. (2011) used a structured process to elicit qualitative information.

Some references use a mix of qualitative and quantitative information to be elicited. The way this information is combined may appear dubious from a probabilistic modelling perspective (e.g., Belkhiria et al., 2018; Soliman et al., 2016; Guinat et al., 2017).

2.1.2 Not documented

Some of the references only mentioned the use of expert elicitation but gave no or very little details about the protocols used, or about the format of the questions (e.g., Panetta, 2011; Wu et al., 2014; Di Fonzo et al., 2016; Hoey et al., 2016). All these can be categorised as unstructured since the process is clearly not thoroughly documented, hence their protocols are not suitable for review.

2.1.3 SEJ mixed with Decision making

Cox et al. (2016) quantified a BN from multiple choice survey data collected on-line, followed by a Delphi-like consensus striving process. Weightings on biosecurity practices were also elicited. This is not good practice given that these weights may mask value judgements and interfere with the probabilistic modelling.

Valdez et al. (2019) used a 10-point scale for experts to assess potential benefits and hazards to be combined into risk measures. Barrett et al. (2009) gave no information about the expert elicitation, but talks of consensus, and point estimates. No formal procedure was describe and the entire process seems like an disorganised structured decision-making exercise. Shortall et al. (2017)

elicited ranks for effectiveness and practicality of measures (without a proper operationalization of these concepts). Oidtmann et al. (2011) asked experts to rank risks, and Guinat et al. (2017) elicited point estimates, calculated semi-quantitative averages, and elicited consequences in the same elicitation.

2.2 Single round protocols

Given that we are proponents of the IDEA protocol, having a single round of estimation, not eliciting bounds and seeking consensus seem sub-optimal, but none of these choices rule such protocols as unstructured.

Lohr et al. (2017) elicited both probabilities, and point estimates for quantities; and sometimes averaged between experts, used a single expert or the consensus of two experts. Brown et al. (2016) used a virtual reality aided elicitation for estimates of the probability of the presence of rock wallabies at given sites (together with an assessment of confidence in the estimates) but had no control group to asses the influence of the virtual reality tool. Pande et al. (2017) used a questionnaire for eliciting pairwise comparisons between the importance of environmental variables. Experts' pairwise answers were combined using the geometric mean and integrated in an analytic hierarchy process matrix.

2.3 Delphi-like protocols for point estimates

Froese et al. (2017) used a Delphi-like approach that started with eliciting partial conditional probability tables (CPTs) and interpolated point estimates to build full CPTs. A second round consisted in giving the full CPTs as feedback to (a subset of) experts, and asking them to modify them. Two different approaches were used: (a) experts specify the weight of each explanatory variable, the overall uncertainty in making this judgement as the variance of a parametric distribution, and the weighted mean function (these are inputs needed in the Agena Software); and (b) elicit only a few rows from the CPT and interpolate.

Murray et al. (2016) elicited CPTs in face-to-face workshops, independently, and then experts discussed the estimates until they achieved group consensus.

In Turbè et al. (2017) a face-to-face workshop was conducted to elicit impact assessments and confidence scores independently. This was followed by a group (consensus) elicitation. For each question, a facilitator presented the distribution of answers from the independent assessments, summarized the available evidence, stimulated discussion and highlighted guidance for scoring impacts and confidence. All experts were then asked to reconsider their answers. Consensus was assumed to be reached when two-thirds of the participants were in agreement. If no consensus was reached the first time, then up to another two voting rounds were conducted, before which the facilitator stimulated further discussion to ensure the discrepancies reflected differences in expert judgement and were not due to overlooking or misinterpretation of evidence, nor to misunderstanding of the scoring rules. This is a very structured protocol, but the operationalisation of the questions asked is not evident.

Eliciting ranks can be a difficult task if value judgements are not separated properly from probabilistic judgements (see discussion above), and when ranks are not properly operationalised. The authors of Gustafson et al. (2018) overcame these difficulties by eliciting relative ranks of risk factors using an estimate-talk-revise protocol (similar to IDEA);

2.4 Delphi-like protocol for uncertainty elicitation

Nicholson and Korb (2017) used no formal protocol to elicit point estimates and some probability intervals, which were further averaged. Both direct elicitation and elicitation aided by an elicitation tool based on verbal clues were used. Apparently "no differences in the time for elicitation or the parameters" was observed. However, the details of these elicitations are unclear.

Kuhnert (2011) presented four case studies, one of which uses a BN elicited from experts. Marginal distributions were elicited from multiple experts during a workshop. The IDEA protocol was both used for probabilities and quantities and one-to-one feedback was provided.

3 Overview

Baxter and Hamilton (2016), Miller et al. (2017), and Vanderhoeven et al. (2017) do not fit in any of the above groupings but they present guidelines for probabilistic risk assessments including SEJ,

or reviews of processes. They contain advice that aligns with the advice previously given in the SEJ literature (e.g., Bolger et al., 2014). One reference presented a software elicitation tool to be used in face-to-face elicitations of distributions through elicitation of percentiles Fisher et al. (2012).

Table 2 places all the reviewed references in different (sometimes overlapping) categories and summarises our findings.

Table 2: Classification of references based on the face-to-face or remote protocols and on the aggregation methods

Protocol and Aggregation	***References***
Face-to-face vs. Remote	Baker et al. (2014); Grigore et al. (2017)
Delphi-like protocol for point estimates — Behavioural aggregation	Gustafson (2010); Morin et al. (2013); Davidson et al. (2013); Fournie et al. (2013); Gustafson et al. (2014); Oidtmann et al. (2014); Soliman et al. (2016); Cox et al. (2016); Brookes and Ward (2017); Turbè et al. (2017)
Delphi-like protocol for point estimates — Mathematical aggregation	Brioudes et al. (2015); Froese et al. (2017); Racicot et al. (2018); Gustafson et al. (2018)
Delphi-like protocol for uncertainty elicitation — Behavioural aggregation	Kuhnert (2011); Carwardine et al. (2012); Singh et al. (2018); Scott et al. (2018); Hemming et al. (2018b)
Delphi-like protocol for uncertainty elicitation — Mathematical aggregation	MA Chades et al. (2015); Murray et al. (2016); Brookes et al. (2017); Rhodes et al. (2017); Robinson et al. (2017); Muellner et al. (2018); Hood et al. (2019); Estevez et al. (2019)
Single round for uncertainty elicitation	Clarke and Jones (2015); Peyre et al. (2016); Suijkerbuijk et al. (2019); Motta et al. (2019)
Single round for point estimates	Soliman et al. (2012); Brookes et al. (2015); Huneau-Salaun et al. (2014); Hoey et al. (2016); Lohr et al. (2017); Carmo et al. (2018)
Point estimates; unknown protocol	Barry and Lin (2010); Jarrad et al. (2011a); Oidtmann et al. (2014); Jarrad et al. (2011b); Wu et al. (2014)
Not an SEJ	Delgado et al. (2016); Lohr et al. (2015); Van Klinken et al. (2016); Leger et al. (2017); Roche et al. (2015); Belkhiria et al. (2018); Soliman et al. (2016); Guinat et al. (2017); Jarrad et al. (2011a); Nicholson and Korb (2017); Valdez et al. (2019)
SEJ mixed with decision-making	Barrett et al. (2009); Shortall et al. (2017); Oidtmann et al. (2011); Valdez et al. (2019); Soliman et al. (2016); Guinat et al. (2017); Rhodes et al. (2017); Schwoerer et al. (2018)
Qualitative	Brouwer et al. (2011); Eggers et al. (2011); Cox et al. (2012); Whittle et al. (2013); Lohr et al. (2015); Roche et al. (2015); Soliman et al. (2016); Leger et al. (2017); Hammer et al. (2019)
Useful summaries, guidelines and software	Fisher et al. (2012); Albert et al. (2012); Baxter and Hamilton (2016); Miller et al. (2017); Vanderhoeven et al. (2017)

A domain independent search would probably generate some (but likely not many) results on remote versus face-to-face elicitations. However, this broader scope was outside the remit of the current project. A quick search identified a couple of studies, one on carbon capture (Baker et al., 2014) and the other on health technology assessment (Grigore et al., 2017), in which small experiments that compared face-to-face and remote elicitation protocols were conducted. Both studies elicited percentiles of distributions of quantities and aggregated the elicited distributions using a equally weighted linear combination. They found that the on-line elicited distributions were less uncertain, but the resulting aggregated distributions were comparable. The authors of Baker et al. (2014) found that even though the elicitation protocol had a significant effect, this was explained (statistically) to some extent by the different (nationalities of) experts.

4 Experimental Design

In the absence of evidence that supports the comparison of face-to-face and remote protocols, we propose an experiment for comparing performance of the approaches used in Australian biosecurity research to answer the question of comparability.

Designing an expert judgement experiment will start with defining expertise and finding experts. It is imperative that results of expert judgement studies include experts (not lay people or students) for critical questions. It is equally important that a sufficient number of diverse[5] experts are recruited to test, validate and repeat experiments. A key criterion to call someone an expert is that they have relevant and demonstrated knowledge related to the topic in question.

What we ask experts and how we formulate the questions are crucial elements of an elicitation exercise, as is the burden we impose on the experts through the elicitation, in terms of time commitment and potential fatigue. From our experience, limiting the number of questions an expert has to answer to no more than 20 in any one experiment, will keep the effects of fatigue low. The format and sequence of questions should be fixed prior to the elicitation, tested in a dry-run for clarity and fairness, and kept constant for all experimental groups.

If one of the treatments involves on-line elicitation (rather than face-to-face), then a dedicated

[5]Some proxies for diversity are gender, age, experience, work place, and background.

website (if available) allows experts to more readily interact and participate. This on-line platform is also important for randomising experimental designs, providing feedback, facilitating discussions.

To be able to empirically evaluate results and make comparisons between treatments, questions should be limited to those for which data can reasonably be obtained to validate judgements in the form they are elicited (i.e., numbers/quantities as realisations of variables whose percentiles were elicited, occurrence of events for elicited probabilities of occurrence). The topics and possible databases from which questions can be developed should be ready before experiments are conducted. Scoring rules for assessing judgements have to be decided upon and used to compare experts' performance in different treatments.

We propose to compare different formats of remote elicitation to face-to-face elicitation and to the hybrid form of the IDEA protocol. Particular aspects of the remote process that we may vary include whether on-line contributions are synchronous versus asynchronous and/or use anonymous versus disclosed identities. To test these different formats, we require multiple replicate groups within each treatment (a minimum of 5 participants per group). It would be ideal to firstly test different remote formats using a Mechanical Turk[6] to ensure that the experiment has sufficient power and on-line capability to detect changes, and refine the requirements accordingly. Such a test is beyond the remit of the current report.

Then the protocol can be tested on expert groups in all three remote, face-to-face, and hybrid settings. Additional experts across multiple workshops are required to complete the experiment.

In the analysis, we would compare the cost, accuracy and calibration of judgements from workshops and remote (or hybrid) elicitations to determine if workshops are cost-effective, and if alternative formats of IDEA are equally effective.

4.1 The frailties of expert judgements

As mentioned in the previous section, a successful experiment requires multiple replicate groups within each treatment. The number of treatments and replicates will influence the reliability of results obtained for comparisons of the technical/operational aspects of elicitations (namely

[6] `https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk`

remote, face-to-face, or hybrid implementation). If the number of replicates is not determined carefully then the results from different treatments may be difficult to interpret, or may be nothing more than noise. Instead of using a huge number of replicates, which would increase the power of our comparisons, we could compromise to a reduced number, given that we have sufficient calibration/validation data.

Using these data we can evaluate experts' accuracy and calibration and assume that good scores correspond to excellent knowledge of the field and the required numerical skills to answer similar questions. These qualities of an expert may be more influential and important for the reliability of elicitation results than the elicitation platform. However, even if that were true, the face-to-face discussion is not only known to boost individual accuracy, but group's accuracy as well. Interactions between experts may work in less evident ways, by some experts asking "all the right questions" and engaging everybody in a conversation which becomes cumbersome on-line.

As a first step towards investigating the above hypothesis, in the next section we propose a modelling exercise which may shed some light on the matter.

4.2 An Easier Alternative

The development of the Risk Return Resource Allocation (RRRA) model involved a considerable number of expert elicitations using at least two different protocols (namely, paper-based questioning and Delphi Cloud). It is highly likely that each of the protocols used calibration questions to develop weights for aggregating the expert opinions. By their nature, calibration questions are a mechanism for determining how well, on average, experts are able to answer questions. If there were a systematic difference in the accuracy of the answers to the calibration questions between the two protocols, then this would be some evidence about the relative quality of the protocols in obtaining accurate answers from experts. Briefly put: we propose to use the calibration questions, which were intended and designed to compare the expert's accuracy, to instead compare the accuracy of the protocols.

Of course, in the absence of a designed experiment there are many other factors that could explain such differences, for example, the calibration questions for one exercise might simply have

been harder than those for the other. It would be necessary to assess the relative difficulty of the questions in order to assuage such concerns. Even if the analysis of the calibration questions were not definitive, regardless the outcomes could guide the experimental design because it would assist in (i) articulating what an important difference in quality looks like, and (ii) providing preliminary estimates for sources of variation that are essential for efficient experimental design.

We recommend that the department undertakes a desktop review of all of the expert elicitation protocols carried out internally, with a particular emphasis on identifying the calibration questions, the experts' answers to the calibration questions, and the actual (true) answers. This resource could then be subjected to a simple regression modelling exercise that assesses the accuracy of the experts' answers to the calibration questions as a function of various factors (depending on the information available), including the SEJ protocol. If the experts can be identified, then more statistical power might be available by including an (anonymised) expert identifier in the statistical model, especially if some experts have been exposed to more than one protocol. A draft candidate model is sketched below, the suitability of which depends entirely on the available information.

$$y_{ijk} = \beta_0 + \beta_1 x_i + b_j + \epsilon_{ijk} \tag{1}$$

where y_{ijk} is the Brier score[7] for calibration question k answered by expert j in study i; x_i is an indicator variable for the SEJ protocol used in study i (e.g., $x_i = 0$ for paper and $x_i = 1$ online); $b_j \sim N(0, \sigma_b^2)$ is an expert-specific random effect; $\epsilon_{ijk} \sim N(0, \sigma^2)$ is a question, expert, study-specific random error; β_0 is a baseline error rate, and β_1 is the estimate of the difference in difficulty of the questions between the protocols identified by x_i.

If this model fits the data acceptably well then a point and interval estimate of β_1 can be considered indicative of the relative difficulty of the calibration questions under each elicitation protocol. If there are no apparent differences between how hard the calibration questions are to answer across the protocols (this consideration will be fuzzy, unfortunately) then we may feel confident that the protocols themselves are operationally comparable in outcome.

[7] see, for example `https://en.wikipedia.org/wiki/Brier_score`

5 Conclusion and Recommendations

After an extensive literature search in the biosecurity domain, the research team has found no evidence that supports the comparison of face-to-face and remote expert elicitation protocols. This suggests the need of reviewing the literature in a domain independent fashion.

It also suggests that the ever changing science of expert elicitation is, relatively speaking, still in its infancy. The need for thorough investigations prior to introducing a new feature to protocols (as is the case with the recently introduced remote variants) is maybe being overwritten by the urgency of the embedding models much needed quantification. However, if such behaviour pertains, there will be no guarantees of repeatability across different novel protocols.

Even though many of the constituting steps of the current protocols have passed the test of time and have been proven beneficial over and over again, such investigations should continue, every time a variation is proposed.

In the meantime, if previously collected expert elicited data can be analysed in a statistical sensible way, this can inform meaningful experimental designs that in turn may answer many o the questions asked in this report.

We recommend that the RRRA team take every pain to try to source calibration questions, expert answers, and true answers from the previous expert elicitation protocols. These results can be used in a statistical model to try to distinguish between the categories, or to assist in designing a statistically valid experiment that may be carried out in future.

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Lowchoy, S., Mengersen, K., Rousseau, J., 2012. Combining expert opinions in prior elicitation. Bayesian Analysis 7, 503–532.

Aspinall, W., 2010. A route to more tractable expert advice. Nature 463, 294–295.

Aspinall, W., Cooke, R., 2013. Quantifying scientific uncertainty from expert judgement elicitation. In: Rougier, J., Sparks, S., Hill, L. (Eds.), Risk and uncertainty assessment for natural hazards. Cambridge University Press, Cambridge, Ch. 10, pp. 64–99.

Baker, E., Bosetti, V., Jenni, K., Ricci, E., 2014. Facing the experts: Survey mode and expert elicitation. Tech. rep., Climate Change and Sustainable Development, Fondazione Eni Enrico Mattei.

Barrett, S., Whittle, P., Mengersen, K., Stoklosa, R., 2009. Biosecurity threats: the design of surveillance systems, based on power and risk. Environmental and Ecological Statistics 17, 503–519.

Barry, S., Lin, X., 2010. Point of truth calibration: Putting science into scoring systems. Tech. rep., Australian Centre of Excellence for Risk Analysis.

Baxter, P., Hamilton, G., 2016. An analysis of future spatiotemporal surveillance for biosecurity, `https://eprints.qut.edu.au/115022/`.

Belkhiria, J., Hijmans, R., Boyce, W., Crossley, B. M., Martínez-Lopez, B., 2018. Identification of high risk areas for avian influenza outbreaks in california using disease distribution models. PLOS ONE 13 (1), 1–15.
URL `https://doi.org/10.1371/journal.pone.0190824`

Bolger, F., Hanea, A., O'Hagan, A., Mosbach-Schulz, O., Oakley, J., Rowe, G., Wenholt, M., 2014. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal 12 (6), 278 pp.

Brioudes, A., Warner, J., Hedlefs, R., Gummow, B., 2015. Diseases of livestock in the pacific islands region: Setting priorities for food animal biosecurity. Acta tropica 143.

Brookes, V., Hernández-Jover, M., Holyoake, P., Ward, M. P., 2015. Industry opinion on the likely routes of introduction of highly pathogenic porcine reproductive and respiratory syndrome into Australia from south-east Asia. Australian Veterinary Journal 93 (1-2), 13–19.

Brookes, V., Keponge-Yombo, A., Thomson, D., Ward, P., 2017. Risk assessment of the entry of canine-rabies into Papua New Guinea via sea and land routes. Preventive Veterinary Medicine 145, 49–66.

Brookes, V., Ward, M., 2017. Expert opinion to identify high-risk entry routes of canine rabies into Papua New Guinea. Zoonoses and Public Health 64 (2), 156–160.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/zph.12284

Brouwer, A., Hill, A., Woodward, M., 2011. What makes a salmonella strain epidemic? an expert opinion workshop. The Veterinary record 168, 1–4.

Brown, R., Bruza, P., Heard, W., Mengersen, K., Murray, J., 2016. On the (virtual) getting of wisdom: immersive 3D interfaces for eliciting spatial information from experts. Spatial Statistics 18, 318–331.
URL https://eprints.qut.edu.au/98473/

Burgman, M., 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.

Carey, J., Burgman, M., 2008. Linguistic uncertainty in qualitative risk analysis and how to minimize it. Annals of the New York Academy of Sciences 1128, 13–17.

Carmo, L., Nielsen, L., Alban, L., da Costa, P., Schupbach-Regula, G., Magouras, I., 2018. Veterinary expert opinion on potential drivers and opportunities for changing antimicrobial usage practices in livestock in Denmark, Portugal, and Switzerland. Frontiers in Veterinary Science 5, 29.

Carwardine, J., O'Connor, T., Legge, S., Mackey, B., Possingham, H. P., Martin, T. G., 2012. Prioritizing threat management for biodiversity conservation. Conservation Letters 5 (3), 196–204.

Chades, I., Nicol, S., van Leeuwen, S., Walters, B., Firn, J., Reeson, A., Martin, T., Carwardine, J., 2015. Benefits of integrating complementarity into priority threat management. Conservation Biology 29 (2), 525–536.

Clarke, S., Jones, S., 2015. Bayesian estimation for diagnostic testing of biosecurity risk material in the absence of a gold standard when test data are incomplete. Journal of Agricultural, Biological, and Environmental Statistics 20 (3), 389–408.
URL https://doi.org/10.1007/s13253-015-0214-5

Clemen, R., Winkler, R., 1999. Combining probability distributions from experts in risk analysis. Risk Analysis 19, 187–203.

Cooke, R., 1991. Experts in uncertainty: Opinion and subjective probability in science. Environmental Ethics and Science Policy Series. Oxford University Press.

Cooke, R., 2015. The aggregation of expert judgment: Do good things come to those who weight? Metron 35, 12–15.

Cooke, R., ElSaadany, S., Huanga, X., 2008. On the performance of social network and likelihood-based expert weighting schemes. Reliability Engineering and System Safety 93 (5), 745–756.

Cooke, R., Goossens, L., 2000. Procedures guide for structural expert judgement in accident consequence modelling. Radiation Protection Dosimetry 90(3), 303–309.

Cooke, R., Goossens, L., 2008. TU Delft expert judgment data base. Reliability Engineering and System Safety 93 (5), 657–674.

Cox, R., Crawford, W. R.and Hurnik, D., Sanchez, J., 2016. Use of Bayesian Belief Network techniques to explore the interaction of biosecurity practices on the probability of porcine disease

occurrence in Canada. Preventive Veterinary Medicine 131, 20–30.
URL http://www.sciencedirect.com/science/article/pii/S0167587716301878

Cox, R., W., C., Sanchez, J., 2012. The use of expert opinion to assess the risk of emergence or re-emergence of infectious diseases in Canada associated with climate change. PLOS ONE 7 (7), 1–13.

Dalkey, N., 1969. An experimental study of group opinion: the Delphi method. Futures 1, 408–426.

Davidson, A., Marnie, L., Hewitt, C., 2013. The role of uncertainty and subjective influences on consequence assessment by aquatic biosecurity experts. Journal of Environmental Management 127, 103–113.

Delgado, J., Pollard, S., Pearn, K., Snary, E., Black, E., Prpich, G., Longhurst, P., 2016. UK foot and mouth disease: A systemic risk assessment of existing controls. Risk Analysis 37, 1768–1782.

Di Fonzo, M., Possingham, H., Probert, W., Bennett, J., Nalini, L. J., Tulloch, A., O'Connor, S., Densem, J., Maloney, R., 2016. Evaluating trade-offs between target persistence levels and numbers of species conserved. Conservation Letters 9, 51–57.

Eggers, S., Verrill, L., Bryant, C., Thorne, S., 2011. Developing consumer-focused risk communication strategies related to food terrorism. International Journal of Food Safety 4, 45–62.

Estevez, R., Mardones, F., Alamos, F., Arriagada, G., Carey, J., Correa, C., Joaquin, E., Gaete, A., Gallardo, A., Ibarra, R., Ortiz, C., Rozas-Serri, M., Sandoval, O., Santana, J., Gelcich, S., 2019. Eliciting expert judgements to estimate risk and protective factors for Piscirickettsiosis in Chilean salmon farming. Aquaculture 507, 402–410.

Ferrell, W., 1994. Discrete subjective probabilities and decision analysis: elicitation, calibration and combination. In: Subjective probability. Vol. Eds. Wright, G. and Ayton, P. Cambridge Press, New York.

Fisher, R., O'Leary, R., Low-Choy, S., Mengersen, K., Caley, M., 2012. A software tool for elici-

tation of expert knowledge about species richness or similar counts. Environmental Modelling & Software 30, 1–14.

Fournie, G., Jones, B., Beauvais, W., Lubroth, J., Njeumi, F., Cameron, A., Pfeiffer, D., 2013. The risk of rinderpest re-introduction in post-eradication era. Preventive Veterinary Medicine 113, 175–184.

French, S., 1985. Group consensus probability distributions: a critical survey. Bayesian Statistics 2. Eds Bernardo J.M., De Groot M.H., Lindley D.V. and Smith A.F.M. North-Holland, Amsterdam: Oxford University Press, 183–201.

Froese, J., Smith, C., Durr, P., McAlpine, C., van Klinken, R., 2017. Modelling seasonal habitat suitability for wide-ranging species: Invasive wild pigs in northern Australia. PLOS ONE 12 (5), 1–22.

Grigore, B., Peters, J., Hyde, C., Stein, K., 2017. EXPLICIT: A feasibility study of remote expert elicitation in health technology assessment. BMC Medical Informatics and Decision Making 17, 1–10.

Guinat, C., Vergne, T., Jurado-Diaz, C., Sánchez-Vizcaíno, J., Dixon, L., Pfeiffer, D., 2017. Effectiveness and practicality of control strategies for African swine fever: what do we really know? Veterinary Record 180 (4), 97–97.
URL `https://veterinaryrecord.bmj.com/content/180/4/97`

Gustafson, L., 2010. Viral hemorrhagic septicemia virus (VHSV IVb) risk factors and association measures derived by expert panel. Preventive veterinary medicine 94, 128–139.

Gustafson, L., Antognoli, M., Lara Fica, M., Ibarra, R., Mancilla, J., Sandoval, O., Enriquez, R., Perez, A., Aguilar, D., Madrid, E., Bustos, P., Clement, A., Godoy, M., Johnson, C., Remmenga, M., 2014. Risk factors perceived predictive of ISA spread in Chile: Applications to decision support. Preventive Veterinary Medicine 117, 276–285.

Gustafson, L., Jones, R., Dufour-Zavala, L., Jensen, E., Malinak, C., McCarter, S., Opengart, K., Quinn, J., Slater, T., Delgado, A., Talbert, M., Garber, L., Remmenga, M., Smeltzer, M., 2018.

Expert elicitation provides a rapid alternative to formal case-control study of an H7N9 Avian Influenza outbreak in the United States. Avian Diseases 62, 201–209.

Hammer, C., Brainard, J., Hunter, P., 2019. Rapid risk assessment for communicable diseases in humanitarian emergencies: validation of a rapid risk assessment tool for communicable disease risk in humanitarian emergencies. Global Biosecurity 1(2).

Hanea, A., Burgman, M., Hemming, V., 2018a. IDEA for Uncertainty Quantification. In: L.C., D., A., M., J., Q. (Eds.), Elicitation: The Science and Art of Structuring Judgement. International Series in Operations Research & Management Science, Springer, Cham, pp. 95–117.

Hanea, A., McBride, M., Burgman, M., Wintle, B., 2018b. The value of performance weights and discussion in aggregated expert judgments. Risk Analysis 38, 1781–1794.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., Mascaro, S., Manning, B., 2016. $I_{nvestigate}D_{iscuss}E_{stimate}A_{ggregate}$ for structured expert judgement. International Journal of Forecasting 33 (1), 267–279.

Hemming, V., Burgman, M. A., Hanea, A., McBride, M., Wintle, B., 2018a. A practical guide to structured expert elicitation using the IDEA protocol. Methods in Ecology and Evolution 9 (1), 169–180.

Hemming, V., M., H., F., J., Rumpff, L., 2018b. NSW Koala research plan: Expert elicitation. Tech. rep., Report prepared by the Centre of Environmental and Economic Research, The University of Melbourne, Australia.

Hoey, J., Campbell, M., Hewitt, C.and Gould, B., Bird, R., 2016. *Acanthaster planci* invasions: Applying biosecurity practices to manage a native boom and bust coral pest in Australia. Management of Biological Invasions 7, 213–220.

Hood, Y., Starkey, C., Sadler, J., Poldy, J., Robinson, A., 2019. Biosecurity system reforms and the development of a risk-based surveillance and pathway analysis system for ornamental fish imported into Australia. Frontiers in Veterinary Science 167, 159–168.

Huneau-Salaun, A., Stark, K., Pereira Mateus, A., Lupo, C., Lindberg, A., Bouquin-Leneveu, S., 2014. Contribution of meat inspection to the surveillance of poultry health and welfare in the European Union. Epidemiology and Infection 143, 1–14.

Jarrad, F., Barrett, S., Murray, J., Parkes, J.and Stoklosa, R., Mengersen, K., Whittle, P., 2011a. Improved design method for biosecurity surveillance and early detection of non-indigenous rats. New Zealand Journal of Ecology 35, 132–144.

Jarrad, F., Barrett, S., Murray, J., Stoklosa, R.and Whittle, P., Mengersen, K., 2011b. Ecological aspects of biosecurity surveillance design for the detection of multiple invasive animal species. Biological Invasions 13, 803–818.

Kuhnert, P., 2011. Four case studies in using expert opinion to inform priors. Environmetrics 22, 662–674.

Leger, A., De Nardi, M., Simons, R., Adkin, A., Ru, G., Estrada-Pena, A., Stark, K., 2017. Assessment of biosecurity and control measures to prevent incursion and to limit spread of emerging transboundary animal diseases in Europe: An expert survey. Vaccine 35 (44), 5956–5966.

Lohr, C., Passeretto, K., Lohr, M., Keighery, G., 2015. Prioritising weed management activities in a data deficient environment: the Pilbara islands, Western Australia. Heliyon 1 (4), e00044.
URL http://www.sciencedirect.com/science/article/pii/S2405844015302747

Lohr, C., Wenger, A., Woodberry, O., Pressey, R., Morris, K., 2017. Predicting island biosecurity risk from introduced fauna using Bayesian Belief Networks. Science of The Total Environment 601–602, 1173–1181.

Miller, J., Burton, K., Fund, J., Self, A., 2017. Process review for development of quantitative risk analyses for transboundary animal disease to pathogen-free territories. BioResearch Open Access 6 (1), 133–140.

Morgan, M., 2015. Our knowledge of the world is often not simple: Policymakers should not duck that fact, but should deal with it. Risk Analysis 35, 19–20.

Morin, L., Heard, T., Scott, J., Sheppard, A., Dhileepan, K., Osunkoya, O., Van Klinken, R., 2013. Prioritization of weed species relevant to Australian livestock industries for biological control. Tech. rep., Meat & Livestock Australia Limited.

Motta, P., Garner, G., Hovari, M., Alexandrov, T., Bulut, A., Fragou, I. A., Sumption, K., 2019. A framework for reviewing livestock disease reporting systems in high-risk areas: assessing performance and perceptions towards foot and mouth disease reporting in the Thrace region of Greece, Bulgaria and Turkey. Transboundary and Emerging Diseases 66 (3), 1268–1279.

Muellner, P., Hodges, D., Ahlstrom, C., Newman, M., Davidson, R., Pfeiffer, D., Marshall, J., Morley, C., 2018. Creating a framework for the prioritization of biosecurity risks to the New Zealand dairy industry. Transboundary and Emerging Diseases 65 (4), 1067–1077.

Murray, J., Jansen, C., De Barro, P., 2016. Risk associated with the release of Wolbachia-infected *Aedes aegypti* mosquitoes into the environment in an effort to control dengue. Frontiers in Public Health 4, 43.
URL `https://www.frontiersin.org/article/10.3389/fpubh.2016.00043`

Nicholson, A., Korb, K., 2017. Bayesian networks for import risk assessment. In: Robinson, A., Walshe, T., Burgman, M., Nunn, M. (Eds.), Invasive Species: Risk Assessment and Management. Cambridge University Pess, Cambridge, pp. 181–205.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities. Wiley, London.

Oidtmann, B., Crane, C., Thrush, M., Hill, B., Peeler, E., 2011. Ranking freshwater fish farms for the risk of pathogen introduction and spread. Preventive Veterinary Medicine 102 (4), 329–340.

Oidtmann, B., Peeler, E., Thrush, M., Cameron, A., Reese, A., Dunn, P., Lyngstad, T., Tavornpanich, S., Brun, E., Stark, K., 2014. Expert consultation on risk factors for introduction of infectious pathogens into fish farms. Preventive Veterinary Medicine 115, 238–254.

Pande, A., González Acosta, H., Brangenberg, N., Knight, B., 2017. A risk-based surveillance design

for the marine pest Mediterranean fanworm *Sabella spallanzanii* (Gmelin, 1791) (Polychaeta: Sabellidae) — a New Zealand case study.

Panetta, D., 2011. Predicting the cost of eradication for 41 Class 1 declared weeds in Queensland. Plant Protection Quarterly 26, 42–46.

Peyre, M., Choisy, M., Sobhy, H., Kilany, W., Gely, M., Tripodi, A., Dauphin, G., Saad, M., Roger, F., Lubroth, J., Makonnen, Y., 2016. Added value of avian influenza (H5) day-old chick vaccination for disease control in Egypt. Avian Diseases 60, 245–252.

Racicot, M., Zanabria, R., Leroux, A., Ng, S., Cormier, M., Tiwari, A., Aklilu, S., Currie, R., Arsenault, J., Griffiths, M., Holley, R., Gill, T., Charlebois, S., Quessy, S., 2018. Quantifying the impact of food safety criteria included in the Canadian Food Inspection Agency risk assessment model for food establishments through expert elicitation. Food Control 92, 450–463.

Radia, D., Bond, K., Limon, G., van Winden, S., Guitian, J., 2013. Relationship between peri-parturient management, prevalence of MAP and preventable economic losses in UK dairy herds. Veterinary Record 173 (14), 343–343.

Rhodes, J., Hood, A., Melzer, A., Mucci, A., 2017. Queensland Koala Expert Panel: A new direction for the conservation of koalas in Queensland. Tech. rep., RA report to the Minister for Environment and Heritage Protection. Queensland Government, Brisbane.

Robinson, A., Brockerhoff, E., Ormsby, M., 2017. Scoping the value and performance of interventions across the NZ Biosecurity system. Tech. rep., Report prepared by the Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne, Australia.

Roche, S., Costard, S., Meers, J., Field, H., Breed, A., 2015. Assessing the risk of Nipah virus establishment in Australian flying-foxes. Epidemiology and Infection 143 (10), 2213–2226.

Rowe, G., Wright, G., 2001. Expert opinions in forecasting: the role of the Delphi technique. In Principles of forecasting: A handbook for researchers and practitioners Norwell: Kluwer Academic Publishers, 125–144.

Schwoerer, T., Little, J., Hayward, G., 2018. Quantifying expert opinion using a discrete choice model: Will invasive *Elodea* spp. threaten wild salmonids in Alaska?, Working Paper.

Scott, A., Toribio, J., Singh, M., Groves, P., Barnes, B., Glass, K., Moloney, B., Black, A., Hernandez-Jover, M., 2018. Low- and high-pathogenic avian influenza H5 and H7 spread risk assessment within and between Australian commercial chicken farms. Frontiers in Veterinary Science 5, 63.

Shortall, O., Green, M., Brennan, M., Wapenaar, W., Kaler, J., 2017. Exploring expert opinion on the practicality and effectiveness of biosecurity measures on dairy farms in the United Kingdom using choice modeling. Journal of Dairy Science 100 (3), 2225–2239.

Singh, M., Toribio, J., Scott, A., Groves, P., Barnes, B., Glass, K., Moloney, B., Black, A., Hernandez-Jover, M., 2018. Assessing the probability of introduction and spread of avian influenza (AI) virus in commercial Australian poultry operations using an expert opinion elicitation. PLOS ONE 13 (3), 1–19.

Soliman, T., Macleod, A., Mumford, J., Nghiem, l., Tan, H., Papworth, S., Corlett, R., Carrasco, L. R., 2016. A regional decision support scheme for pest risk analysis in Southeast Asia. Risk analysis 36, 904–913.

Soliman, T., Mourits, M., Lansink, A. O., van der Werf, W., 2012. Quantitative economic impact assessment of an invasive plant disease under uncertainty — A case study for potato spindle tuber viroid (PSTVd) invasion into the European Union. Crop Protection 40, 28–35.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., Burgman, M., 2010. Reducing overconfidence in the interval judgments of experts. Risk Analysis 30, 512–523.

Suijkerbuijk, A., Over, E., Opsteegh, M., Deng, H., Gils, P. v., Bonacic M., A. A., Lambooij, M., Polder, J., Feenstra, T., Giessen, J. v. d., Wit, G. d., Mangen, M. J., 2019. A social cost-benefit analysis of two One-Health interventions to prevent toxoplasmosis. PLOS ONE 14, 1–16.

Turbè, A., Strubbe, D., Mori, E., Carrete, M., Chiron, F., Clergeau, P., Le Louarn, M., Luna, A., Menchetti, M., Nentwig, W., Pârâu, L., Postigo, J., Rabitsch, W., Senar, J. C., Tollington, S.,

Vanderhoeven, S., Weiserbs, A., Shwartz, A., 2017. Assessing the assessments: evaluation of four impact assessment protocols for invasive alien species. Diversity and Distributions 23, 1–11.

Valdez, R., Kuzma, J., Cummings, C., Peterson, M., 2019. Anticipating risks, governance needs, and public perceptions of de-extinction. Journal of Responsible Innovation 6 (2), 211–231.

Van Klinken, R., Morin, L., Sheppard, A., Raghu, S., 2016. Experts know more than just facts: eliciting functional understanding to help prioritise weed biological control targets. Biological Invasions 18 (10), 2853–2870.

Vanderhoeven, S., Branquart, E., Casaer, J., Dh́ondt, B., Hulme, P., Shwartz, A., Strubbe, D., Turbé., A., Verreycken, H., Adriaens, T., 2017. Beyond protocols: improving the reliability of expert-based risk analysis underpinning invasive species policies. Biological Invasions 19, 2507–2517.

von der Gracht, H., 2012. Consensus measurement in Delphi studies: Review and implications for future quality assurance. Technological Forecasting & Social Change 79, 1525–1536.

Whittle, P., Stoklosa, R., Barrett, S., Jarrad, F., Majer, J., Martin, P., Mengersen, K., 2013. A method for designing complex biosecurity surveillance systems: Detecting non-indigenous species of invertebrates on barrow island. Diversity and Distributions 19, 629–639.

Woudenberg, F., 1991. An evaluation of Delphi. Technological Forecasting and Social Change 40, 131–150.

Wu, P., Pitchforth, J., Mengersen, K., 2014. A hybrid queue-based Bayesian Network framework for passenger facilitation modelling. Transportation Research Part C: Emerging Technologies 46, 247–260.