Report Cover Page

ACERA Project

607

Title

Evaluation and Development of Formal Consensus Methods

Author(s) / Address (es)

Mark Colyvan, Mark Burgman, Aidan Lyon, Helen Regan, Katie Steele

Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)

Final project report

Summary

The objectives of the project were to survey, test, and develop formal consensus methods in the existing literature, consider the properties of these methods from a more pragmatic standpoint than is usually considered, and determine where the methods fit into larger decision–making problems. The aim was to develop guidelines for choosing a formal model that offers a repeatable, transparent, reliable and understandable basis for resolving group disagreement over various issues in a larger decision problem in bio–security settings.

The main findings from the project are summarised in the ACERA technical report Steele *et al.* 2008, in the Appendix. This document provides a systematic summary and evaluation of various formal methods of consensus and related social choice methods. The literature survey demonstrates how formal group judgement aggregation and consensus methods can be employed to help settle distinct components of a decision problem.

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

Evaluation and Development of Formal Consensus Methods

ACERA Project No.

Mark Colyvan, University of Sydney

Helen Regan, University of California, Riverside

Final project report

May 2009

Acknowledgements

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

Disclaimer

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

Table of contents

List of Figures

1. Executive Summary

The objectives of the project were to survey, test, and further develop formal consensus methods in the existing literature, consider the properties of these methods from a more pragmatic standpoint than is usually considered, and determine where the methods fit into larger decision–making problems. The aim was to develop guidelines for choosing a formal model that offers a repeatable, transparent, reliable and understandable basis for resolving group disagreement over various issues in a larger decision problem in bio–security settings. The specific objectives of the project were to:

1. Survey existing formal consensus methods in the literature.
2. Provide a written report on these methods.
3. Test these methods for suitability of implementation.
4. Describe the strengths and weaknesses of alternatives that represent a range of circumstances under which it is envisioned that such methods will be used.
5. Develop these existing methods and (if appropriate) develop new methods.
6. Write scholarly and more accessible summaries on the findings of the testing and development phase of the project.
7. Provide guidelines and advice on consensus methods, including training experts in the use of these methods.

The project has proceeded well and all goals and objectives have been met in the specified time frame. There have been a number of new findings and implementations of those findings, and these have been presented at a numbers of scientific and other forums. There have been many publications arising from the research work.

The main findings from the project are summarised in the ACERA technical report Steele *et al.* 2008, in the Appendix. This document provides a systematic summary and evaluation of various formal methods of consensus and related social choice methods. The literature survey demonstrates how formal group judgement aggregation and consensus methods can be employed to help settle distinct components of a decision problem.

The technical report analyses the general problem of resolving group disagreement in terms of some key features any particular such problem may have, using a bio–security group decision problem as a case study. These key features are:

1) Whether there is past performance data available for each member of the group.
2) If such data exists, what is the best way to use it.
3) Whether there is a danger of strategic play in the process of resolving the disagreement.
4) Whether there is a general agreement on the equal expertise of the members of the group.

The technical report offers the following flow–chart based on these key features as a method for resolving group disagreement:

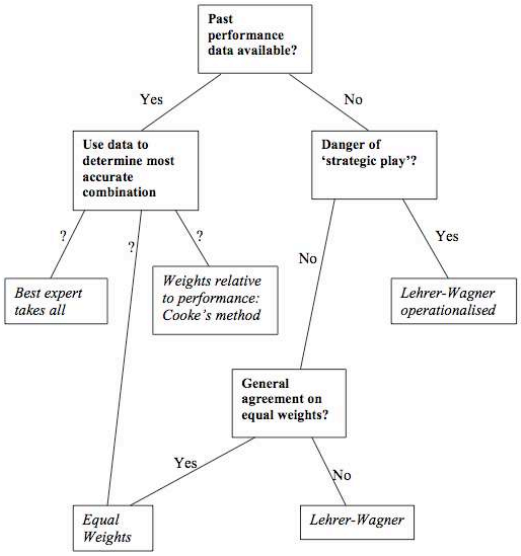


Figure 1.

2. Introduction

Disagreement is a fact a life. Even experts disagree about various scientific matters. Such disagreement can present problems for management and policy decisions when action needs to be taken based on expert judgements. This project investigates and develops various formal methods for settling such disagreements, in which all the available information is utilised.

Disagreement among experts can occur for a number of reasons (Burgman 2005), including differences in expertise, difference in data available to each expert, differences in interpretation of the data, different theoretical biases, different awareness and appreciation of the various uncertainties involved (Regan *et al*. 2002), and different social and political agendas. Although in many cases of interest there may be only one expert, still there can be disagreement when the expert in question changes his or her mind. More recent opinions are not always better than earlier ones. Even in cases of a single expert we may need to address issues of internal disagreement over time and consider ways of aggregating these different opinions.

Settling disagreements among experts has been achieved through various voting methods. However, different voting methods can lead to conflicting results. Despite the predominance of simple majority rule in many scenarios where a disagreement is settled by a "straw poll", there are many voting systems and they deliver different results (Saari 2001). Here is a simple example to illustrate the point. Suppose that we have three potential invasion pathways: 1, 2, and 3. And suppose we have seven experts. Suppose that three of the experts determine that path 1 is the most serious threat, 2 somewhat less serious and path 3 the least serious threat. Suppose that two experts determines that path 2 is the most serious threat, 1 somewhat less serious and 3 the least serious threat. Suppose that the final two experts determine that path 3 is the most serious threat, 2 somewhat less serious and 1 the least serious. (Illustrated in Figure 1.)

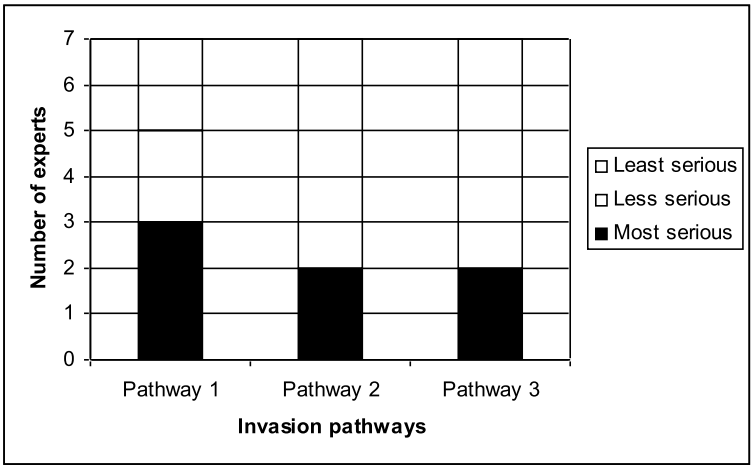


Figure 2.

Under simple *plurality voting*, path 1 would seem to be the favoured choice as the most serious threat, since it receives more votes from experts than any other as the most serious threat. But now consider a *preferential voting* system. Notice that the majority of experts see path 2 as a more serious threat than path 1. That consideration alone would seem to rule path 1 out as the group's considered choice of the most serious threat. It is not clear that one of these two voting systems is fairer, more objective, or more reliable than the other. Things get even more complicated if we introduce other forms of voting such as *weighted preferential voting*, or *approval voting*. The bottom line is that different, apparently fair, voting systems deliver different results. This is known as *the voting paradox*. There is no single voting system that satisfies some very reasonable constraints about what we expect from a such a system (Arrow 1951).

Some voting methods do not attempt to move experts to change their views. The experts hold their own views fixed and the voting system tries to deliver a group preference, and the latter may or may not coincide with the preferences of the members of the group. In any case, the group members hold their own preferences fixed. This is thus a *compromise method*, since the group outcome is typically not the unanimous opinion of the experts. Contrast such methods with a jury, where unanimous approval is required from all jury members. Such methods require the experts in question

to modify their views such that consensus is reached and all participants are satisfied/equally represented.

In light of the well-known problems with voting methods (such as the problem just sketched), it is worth exploring different kinds of group decision-making methods, namely consensus methods. One such method has been applied to conservation management decisions (Regan *et al*. 2006). This method requires experts to provide both their own judgements on the matter at hand (relative seriousness of potential risk pathways, in the example just discussed) as well as judgements of the expertise of others in the group. This is enough to motivate changes in expert opinions in a way that can be formally modelled using simple linear algebra. Moreover, the formal model in question guarantees a consensus outcome in any case where very reasonable conditions are satisfied (Lehrer and Wagner 1981). There are a variety of formal methods for delivering consensus (Steele *et al*. 2008) but they all enjoy the virtue of avoiding the arbitrariness of voting methods. In many cases—especially those where voting methods are seen to break down, as above—formal consensus methods provide an effective and workable alternative.

The objectives of the project were to survey, test, and further develop formal consensus methods in the existing literature, consider the properties of these methods from a more pragmatic standpoint than is usually considered, and determine where the methods may fit in biosecurity decision–making. The aim was to develop guidelines for choosing a formal model that offers a repeatable, transparent, reliable and understandable basis for resolving group disagreement over various issues in a larger decision problem in bio–security settings. The following were the specific objectives of the project:

1. Survey existing formal consensus methods in the literature.
2. Provide a written report on these methods.
3. Test these methods for suitability of implementation.
4. Describe the strengths and weaknesses of alternatives that represent a range of circumstances under which it is envisioned that such methods will be used.
5. Develop these existing methods and (if appropriate) develop new methods.
6. Write scholarly and more accessible summaries on the findings of the testing and development phase of the project.
7. Provide guidelines and advice on consensus methods, including training experts in the use of these methods.

The project results are summarised with respect to these specific objectives in section 4 of this report.

3. Methodology

A number of formal models existing in the judgement consensus and aggregation literature were identified as potentially useful tools for solving real–world, group–decision problems. These models were compared to each other with respect to their mathematical properties, philosophical justification, and their domains of application.

The survey of existing models proceeded by an extensive literature search across the many disciplines that contribute to the development of formal consensus methods (including economics, game and social choice theory, philosophy, mathematics, political and social sciences and computing). Each model was evaluated via two quite different methods: (i) conceptual (ii) and pragmatic. The first consisted of scrutinising each model in terms of its conceptual, philosophical and mathematical coherence, the underlying motivation for the model and its capability of delivering plausible results in well-understood cases. The second was to evaluate whether each model is practical, in the sense that it is implementable and useful to the end users. This required discussions with relevant personnel within the Department of Agriculture, Fisheries and Forestry (DAFF), and others elsewhere, who are seen as possible end users of the methods in question.

Materials necessary for the implementation of a selection of these models were developed. This was done using the computational software program, *Mathematica*.

4. Project Results

The following is a summary of the project's results framed in terms of the project's specific objectives.

4.1. Objective 1: Survey existing formal consensus methods in the literature.

The technical report, Steele *et al* (2008), attached in the Appendix, provides an in-depth, extensive survey and analysis of the formal consensus methods in the literature. Formal consensus methods were reviewed for: a single expert gathering evidence; estimates based on single data source; evidence from a number of data sources; and evidence from a number of experts.

Other surveys were produced during the period of the project. Steele *et al.* (forthcoming) evaluates the "sensitivity" of results of the weighted linear average algorithm used in a class of multi–criteria methods commonly used in environmental decision–making. The report concludes that it is possible to change the final ranking of options by re–calibrating the scoring scales for the criteria. Steele (2007a) reviews a recent monograph—Sarkar (2007)—on scientific group rationality.

4.2. Objective 2: Provide a written report on these methods.

The provided written report on the existing formal consensus methods is the technical report, Steele *et al* (2008), attached in the Appendix.

4.3. Objective 3: Test these methods for suitability of implementation.

The following summarises two ways in which the methods have been tested. For a more extensive discussion, see the technical report in the Appendix.

One way in which the methods were tested for pragmatic suitability was through their implementation in solving real group–decision problems. It was discovered through trying to implement these methods that it can be difficult or impossible to determine certain parameters that the mathematical models of the methods required. Discussion of this issue can be found in Regan et al. (2006). In response to this difficulty discovered through the testing process, a new method for determining these parameters was developed (see section 4.5 below).

One way in which the methods were tested for conceptual suitability was through consideration of the virtues they are intended to have. In any group–decision problem, there two possible virtues (not mutually exclusive or exhaustive) that a solution can have: the solution is a decision that all group members are happy with, even though it may be the "wrong" decision, the solution is the "right" decision, even though all or a substantial proportion of the group members are unhappy with it. Discussion of this issue can be found in Steele *et al.* (2007).

One way in which the methods were tested simultaneously for their conceptual and pragmatic suitability was through computer simulations. One finding from the simulations was that the mathematical models of a given consensus method can be represented in different ways, and some of those representations can be more conducive to making the methods accessible than others. The details of the simulations are reported here as they have not been published elsewhere and are not summarised in the technical report.

For example, the Lehrer–Wagner models takes as inputs the group members individual judgments, p_i, and a weight of opinion assigned to every member by i by every other member j, w_{ij}, and outputs a single judgement, p. One way of thinking about how the model works is that it converts the weights for each member i into an overall weight for that member, and then uses that overall weight in a simple weighted–averaging formula. The overall weights, w_i, are determined in the following way:

$$\begin{pmatrix} w_1 & w_2 & w_3 & \dots & w_n \\ w_1 & w_2 & w_3 & \dots & w_n \\ . & . & . & . & . \\ . & . & . & . & . \\ w_1 & w_2 & w_3 & \dots & w_n \end{pmatrix} = \lim_{m \longrightarrow \infty} \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ w_{n1} & w_{n2} & w_{n3} & \dots & w_{nn} \end{pmatrix}^m$$

where *m* is a non–negative integer. In this way of looking at the model, we can think of the group members' opinions of each other evolving over time, and when those opinions have reached a consensus, the opinion weights are used in a simple weighted–average formula.

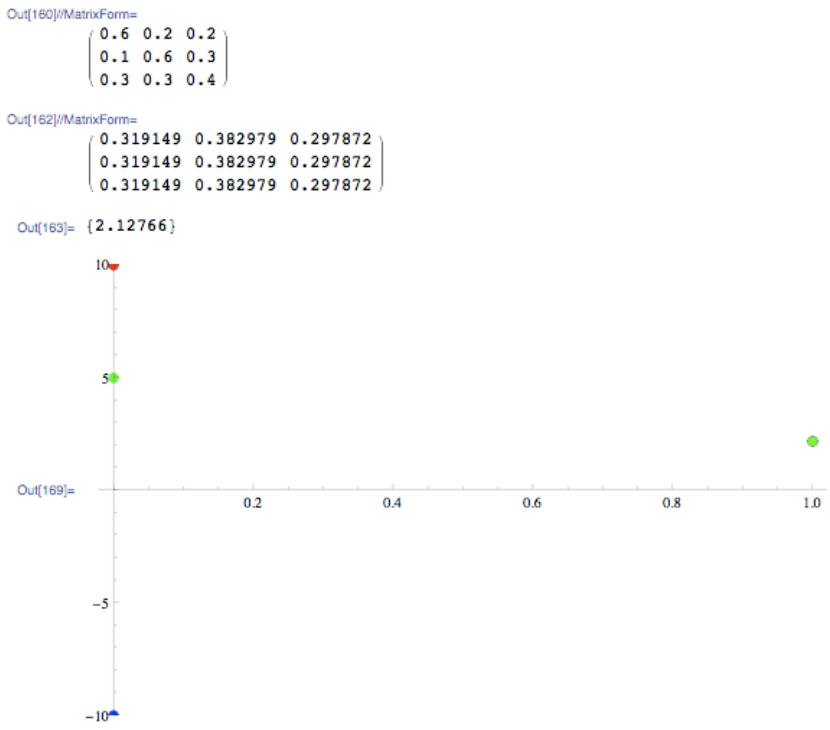
An alternative way of thinking about the model is that judgements (rather than opinions of other experts) evolve over time instead. This is done by taking an initial vector *p* that contains the group's initial judgments and then multiplying that vector by a matrix that contains the individual weights repeatedly until the judgments reach a consensus:

$$\begin{pmatrix} p \\ p \\ . \\ . \\ p \end{pmatrix} = \lim_{m \longrightarrow \infty} \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ w_{n1} & w_{n2} & w_{n3} & \dots & w_{nn} \end{pmatrix}^{m} \begin{pmatrix} p_1 \\ p_2 \\ . \\ . \\ p_n \end{pmatrix}$$

These two alternative ways of thinking about the can be represented graphically in the following way:

Representation 1:

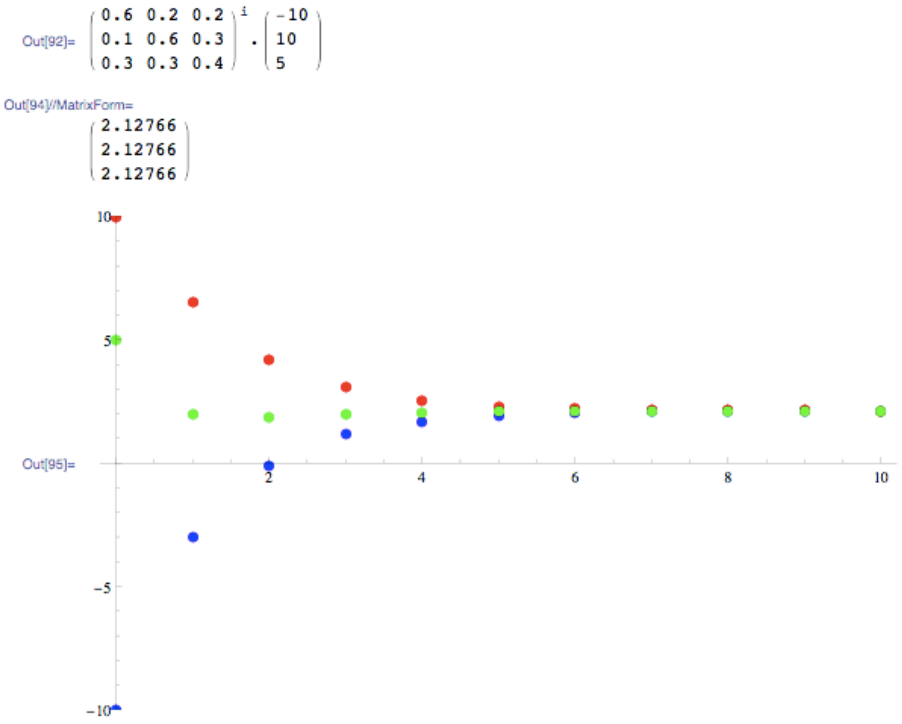
Initially, varying expertise opinions are converted into overall expertise weights, and used in a simple weighted–averaging formula to determine the final judgement value.



In the figure, the three coloured dots represent the judgments made by the group members concerning the quantity of interest.

Representation 2:

Initially, varying judgments evolve as the matrix of opinion weights are repeatedly applied to the judgment vector:



In this second way of representing the model, the final judgment can appear less mysterious to the group, as the trajectories of the individual judgements can be seen to provide an explanation for the final consensus value.

4.4. Objective 4: Describe the strengths and weaknesses of alternatives that represent a range of circumstances under which it is envisioned that such methods will be used.

The strengths and weakness of alternative formal consensus models are discuss in extensive detail in the technical report in the Appendix (pp. 24–34). The report also gives a very accessible detailed analysis of the domains of application of each of the methods, thus detailing the range of circumstances in which the models should be used.

4.5. Objective 5: Develop these existing methods and (if appropriate) develop new methods.

The following is a summary of the development of new and existing methods for formal consensus conducted in the course of this project.

4.5.1. A new method for determination of weights–of–expertise parameters required by formal consensus models.

In response to the problem regarding the determination of parameters that the various formal consensus models require (mentioned in section 4.3), a new method for determining these parameters was developed. This method uses the group members' judgements as inputs and uses a metric function to determine group expert weights (the parameters in question). Letting p_i represent member i's judgment, the expert weight assigned by group member j to group member i is:

$$w_{ij} = \frac{1 - |p_i - p_j|}{\sum_{j=1}^{n} 1 - |p_i - p_j|}$$

One limitation of this method is that it requires the quantity of interest to be non–negative and normalised. This limitation can be problematic in cases where the quantity of interest can take on negative values or where there is disagreement or uncertainty concerning the range of possible values the quantity can take. However, for quantities like probabilities, this new method has many desirable features. Further discussion of this issue can be found in Regan *et al.* (2006).

4.5.2. The development of a new decision–making framework that encapsulates ethical considerations.

Standard decision making frameworks typically ignore ethical considerations in their determination of the optimal decision to make. This is a rather severe limitation in the context of public decision making and bio–security problem solving. In this regard, a new decision–making framework has been developed to account for various ethical considerations. Further details on this new framework can be found in Colyvan et al. (forthcoming a).

4.5.3. New alternative formal theories of risk management and decision making.

Standard formal theories of risk management—such as Kolmogorov's probability axioms and expected utility theory—assume classical sentential logic or set theory. This assumption can render these theories unusable in real–world decision problems that call for a more nuanced approach to questions of logic. This important point has gone largely unnoticed by researchers in the field. Colyvan (2008a) discusses this issue, and considers various ways of dealing with it.

4.5.4. A New Approach to Public Decision Making

Psychological studies have shown that decisions made by humans can systematically diverge from the standard decision–making frameworks (such as expected utility theory). This makes the standard frameworks poor descriptive and predictive models—and arguably, poor normative ones too. To solve these problems, a new approach to public decision–making has been considered in Steele (2006).

4.6. Objective 6: Write scholarly and more accessible summaries on the findings of the testing and development phase of the project.

The technical report in the Appendix is the primary document that addresses this objective. Several other publications also give scholarly and more accessible summaries of the findings of the testing and development phase of the project. These include Colyvan and Regan (2007), Colyvan and Steele (forthcoming), Steele (MSa).

In addition to these publications and the technical report, several international, interdisciplinary conference and workshop presentations were given in an outreach effort to make the formal consensus methods more accessible to general audiences. These presentations included:

- 7 May 2007: Colyvan, M. 'Formal Models of Consensus', presented in the workshop on "Methodological Problems in the Social Sciences" in the Tilburg Center for Logic and Philosophy of Science, Tilburg University, Tilburg, Netherlands.
- March 2006: Regan, H. 'A formal model for consensus and negotiation in land-use planning', presented at the 21st Annual Symposium of the United States Regional Chapter of the International Association for Landscape Ecology, San Diego, USA.
- 9 May 2007: Steele, K. 'Group decision models: Balancing truth and fairness', presented at the AEDA Conservation Planning Workshop, Bardon Conference Centre, Brisbane, Australia.

4.7. Objective 7: Provide guidelines and advice on consensus methods, including training experts in the use of these methods.

All relevant information to prepare guidelines and advice on consensus methods is given in the technical report in the Appendix. In addition to this report, a DAFF consulting session and workshop was held:

- ACERA Workshop on Formal Consensus Methods Aug 21–24 in Sydney (project members plus Melbourne PhD student Marissa McBride) including an afternoon session 2–5pm on 24th August for selected DAFF invitees (*and preliminary follow up conducted with DAFF invitee Roberta Rosselly)*.

And the following joint ACERA and DAFF meeting was held at the University of Melbourne:

- ACERA “Demonstration Project” Meeting (including Mark Burgman and Mark Colyvan along with other members of ACERA and DAFF) 8 Dec 2008 in Melbourne

Future software training sessions are also envisioned in order to train experts and decision makers on how to use the computational software implementation of the consensus methods.

6. References

- Arrow, K.J. 1951. *Social Choice and Individual Values*, New York: Wiley.
- Burgman, M.A. 2005. *Risks and Decisions for Conservation and Environmental Management*, Cambridge: Cambridge University Press, 2005.
- Colyvan, M. 2008a. 'Is Probability the Only Coherent Approach to Uncertainty, *Risk Analysis*, 28(3): 645–652.
- Colyvan, M. 2008b. 'Population Ecology', chapter in S. Sarkar and A. Plutynski (eds.), *A Companion to the Philosophy of Biology*, Blackwell, 301–320.
- Colyvan, M., Cox, D. and Steele, K. forthcoming a. 'Modelling the Moral Dimension of Decisions', *Noûs*.
- Colyvan, M. and Regan, H.M. 2007. 'Legal Decisions and the Reference-Class Problem', *International Journal of Evidence and Proof*, 11(4): 274–285.
- Colyvan, M. and Steele, K. forthcoming. 'Environmental Ethics and Decision Theory: Fellow Travellers or Bitter Enemies?', invited contribution to B. Brown, K. de Laplante, and K. Peacock (eds.), *Handbook of the Philosophy of Science, Volume 11: Philosophy of Ecology*, North Holland/Elsevier.
- Justus, J., Colyvan, M., Regan, H.M. and Maguire, L.A. 2009. 'Buying Into Conservation: Intrinsic Versus Instrumental Value', *Trends in Ecology and Evolution*, forthcoming.
- Lehrer, K. and Wagner, C. 1981. *Rational Consensus in Science and Society*, Dordrecht: D. Reidel Publishing.
- Regan. H.M., Colyvan, M. and Markovchick-Nicholls, L. 2006. 'A Formal Model for Consensus and Negotiation in Environmental Management', *Journal of Environmental Management*, 80(2): 167–176.
- Regan, H.M., Davis, F.W., Andelman, S.J., Widyanata, A. and Freese, M. 2007. 'Comprehensive Criteria for Biodiversity Evaluation in Conservation Planning', *Biodiversity and Conservation*, 16(9): 2715–2728.
- Saari, D.G. 2001. *Decisions and Elections: Explaining the Unexpected*. Cambridge: Cambridge University Press.
- Sarkar, H. 2007. *Group Rationality in Scientific Research*. Cambridge University Press.
- Steele, K. forthcoming. 'Review of Dan Egonsson's *Preference and Information*', *Economics and Philosophy*.
- Steele, K. 2007a. 'Review of Husain Sarkar's *Group Rationality in Scientific Research*', *Notre Dame Philosophical Reviews*, URL= <http://ndpr.nd.edu/review.cfm?id=11484>.
- Steele, K. 2007. 'Planned Changes in Desire', *Proceedings of the Workshop on Logic, Rationality and Interaction*, Beijing: College Publications.
- Steele, K. 2007. 'Distinguishing Indeterminate Belief from "Risk-averse" Preferences', *Synthese*, 158(2): 189–205.
- Steele, K. 2006. 'The Precautionary Principle: A New Approach to Public Decision-Making?', *Law, Probability and Risk*, 5(1):19–31.
- Steele, K., Carmel, Y., Cross, J., and Wilcox, C. forthcoming. 'Uses and Misuses of Multicriteria Decision Analysis (MCDA) in Environmental Decision-Making', *Risk Analysis*.
- Steele, K., Colyvan, M., Regan, H.M. and Burgman, M.A. 2008. 'Survey of Group Consensus Methods', ACERA technical report prepared for the Department of Agriculture Fisheries and Forestry, Australian Centre of Excellence for Risk Analysis Report, 41 pp.
- Steele, K. Regan, H.M. and Colyvan, M. 2007. 'Right Decisions or Happy Decision Makers?', *Social Epistemology*, 21(4): 349–368.
- Wooldridge, M. 2008. 'Qualitative Risk Assessment' in D.W. Schaffner (ed.), *Microbial Risk Analysis of Foods*, Washington DC: ASM Press: 1–28.

***Works under Consideration*:**

- Colyvan, M., Linquist, S. Grey, W., Griffiths, P. Odenbaugh, J. and Possingham, H. MS. 'Current Trends and Future Directions in the Philosophy of Ecology' under consideration at *Ecology and Society*.
- Steele, K. MSa 'Assessing decision rules in the sequential-choice setting', under consideration at *Theory and Decision*.
- Steele, K. MSb 'First- and second-order desire in Jeffrey's decision model', under consideration at *Erkenntnis*.
- Steele, K. MSc 'Reconciling standard "one-shot-only" decision theory with sequential choice', under consideration at *Mind*.

***Works in Progress*:**

- Colyvan, M. 'The Environment is Valuable But Not Infinitely Valuable'
- Colyvan, M., Regan, H.M. and Justus, J. 'The Conservation Game: Survey of Game-theoretic Methods in Conservation-management Applications'.
- Maguire, L.A., Regan, H.M., Martin, T. and Colyvan, M. 'Social Choice Methods in Conservation Management'.
- Steele, K., Colyvan, M., Regan, H.M. and McBride, M. 'A Survey of Formal Consensus Methods for Conservation Management'.
- Steele, K. 'Pareto-optimality and the distinction between deliberation and aggregation in group choice'.

Presentations:

Mark Colyvan's Relevant Presentations:

- 28 May 2009: 'Consensus, Compromise, and the "Wisdom" of Crowds', presented to the Department of Philosophy, University of Melbourne, Melbourne, Australia.
- 30 September 2008: 'Consensus Among Experts', presented at the Society for Risk Analysis Conference at ANU, Canberra, Australia.
- 29 September 2008: 'Aggregating Beliefs: Consensus versus Compromise', presented at the Society for Risk Analysis workshop "Philosophical Reflections on Risk Management and Complex Systems" at ANU, Canberra, Australia.
- 10 April 2008: 'What's Maths Got to Do with It?', presented at the Sydney-Tilburg Reduction and the Special Sciences conference, Tilburg University, Tilburg, The Netherlands.
- 7 May 2007: 'Formal Models of Consensus', presented in the workshop on "Methodological Problems in the Social Sciences" in the Tilburg Center for Logic and Philosophy of Science, Tilburg University, Tilburg, Netherlands.
- 3 May 2007: 'Probability and Law', presented to the Departments of Theoretical and Practical Philosophy at Lund University, Lund, Sweden.
- 23 April 2007: 'Relative Expectation Theory', presented to the Department of Philosophy at the Ludwig-Maximilians University, Munich, Germany.
- 16 December 2006: 'The Use of Mathematics to Describe Biological Systems', presented at the 3rd Queensland Biohumanities Conference, University of Queensland, Brisbane, Australia.
- 9 October 2006: 'Do the Laws of Ecology Lie?' presented in the History and Philosophy of Science seminar series at the University of Sydney, Sydney, Australia.
- 25 August 2006: 'Probability in Law' presented to the TC Beirne School of Law at the University of Queensland, Brisbane, Australia.
- 18 July 2006: 'Risk and the Limitations of (Classical) Probability Theory' presented at the Australian Society for Risk Analysis Conference at the University of Melbourne, Melbourne, Australia.

- 2 July 2006: 'What's Maths Got to Do with It?', Presidential Address at the 2006 Australasian Association of Philosophy Conference at the Australian National University, Canberra, Australia.
- 29 June 2006: 'Mathematical Models in Ecology and Conservation Biology', presented at the 2nd Queensland Biohumanities Conference, University of Queensland, Brisbane, Australia.

Helen Regan's Relevant Presentations:

- March 2006: 'A formal model for consensus and negotiation in land-use planning', presented at the 21st Annual Symposium of the United States Regional Chapter of the International Association for Landscape Ecology, San Diego, USA.

Katie Steele's Relevant Presentations:

- 8 July 2008: 'Bayesian Reasoning and the Scope of IBE', presented at the Australasian Association of Philosophy Conference at La Trobe University, Melbourne.
- 3 April 2008: 'Conflicts Between Truth-Tracking and Fairness-Based Ideals for Aggregating Group Judgments', presented at the Tilburg University Centre for Logic and Philosophy of Science.
- 5–9 August 2007: 'Planned Preference Change', presented at the Workshop on Logic, Rationality and Interaction (LORI), Beijing, China.
- 25 July 2007: 'Right Decisions or Happy Decision-Makers', presented to the 'Choice Group' at the London School of Economics, London, UK.
- 5–6 July 2007: 'Planned Preference Change', presented at the British Society for the Philosophy of Science Annual Meeting, Bristol, UK.
- 9 May 2007: 'Group decision models: Balancing truth and fairness', presented at the AEDA Conservation Planning Workshop, Bardon Conference Centre, Brisbane, Australia.
- 22 November 2006: Comments on Philip Pettit's 'Rationality, reasoning and regulation: the case of group agents', Centre for Time "Minds, Mobs and Memories" Conference, University of Sydney.
- 26 May 2006: 'What is it Rational to Value?', presented at the Formal Epistemology Workshop, UC Berkeley.
- 15 April 2006: 'Modelling the Moral Dimension of Decisions', 8th Annual CMU/Pitt Graduate Student Philosophy Conference, Pittsburgh.

***Conferences and Workshops*:**

- ACERA "Demonstration Project" Meeting (including Mark Burgman, and Mark Colyvan along with other members of ACERA and DAFF) 8 Dec 2008 in Melbourne
- NCEAS Working group Part 4 (including Mark Burgman, Helen Regan and Mark Colyvan) 15–19 Nov 2008 in Santa Barbara
- NCEAS Working group Part 3 (including Mark Burgman, Helen Regan and Mark Colyvan) 9–13 June 2008 in Santa Barbara
- ACERA workshop on The Value of Field Research in Ecology (attended by Katie Steele and Mark Burgman) 11–12 March 2008.
- NCEAS Working group Part 2 (including Mark Burgman, Helen Regan and Mark Colyvan) 3–7 Dec 2007 in Santa Barbara
- ACERA Workshop on Formal Consensus Methods Aug 21–24 in Sydney (project members plus Melbourne PhD student Marissa McBride) including an afternoon session 2–5pm on 24th August for selected DAFF invitees (*and preliminary follow up conducted with DAFF invitee Roberta Rosselly)*.
- ACERA Robust Multi-Criteria Decision Analysis Workshop (Helen Regan, Mark Burgman and Katie Steele attended), 12–17 Aug 2007 in Hobart.

- AEDA Conservation Planning Workshop (Katie Steele attended), 8–11 May in Brisbane
- NCEAS Working group Part 1 (including Mark Burgman, Helen Regan and Mark Colyvan) Jan–Feb 2007 in Santa Barbara

7. Appendix

Steele *et al*. (2008) 'Survey of 'Group' Consensus Methods' attached.

Survey of group ‘consensus’ methods

Australian Centre of Excellence for Risk Analysis

Project: Evaluation and development of formal consensus methods

Katie Steele
Mark Colyvan
Helen Regan
Mark Burgman

Contents

1 Executive Summary

1.1 Aim

The objectives of this project are to: review formal 'consensus' methods in the existing literature; consider the properties of these methods from a more pragmatic standpoint than is usually considered; and determine where the methods fit into larger decision-making problems, using a pertinent problem in plant pest management as a case study. The idea is to develop guidelines for choosing a formal model that offers a repeatable, transparent, reliable and understandable basis for resolving group disagreement over various issues in a larger decision problem in biosecurity settings.

1.2 Summary

Group decision-making is notoriously difficult. Discussions can be longwinded and irrelevant, some group members' expertise may not be well utilised, and there may be hidden agendas and bullying within the group, especially when the issues are complex and there are important policy implications at stake. Formal 'consensus' or group aggregation methods provide structure to decision processes, with the objective that they are more efficient, effective, transparent and fair.

To properly examine the role of formal group aggregation methods, it is useful to consider them within the context of a broader decision problem. This report focuses on a decision model that was developed in ACERA Project 0707 to assist in the prioritising of non-indigenous non-primary industry pest threats. The overall task is to score and subsequently categorise pests in terms of expected impact. This is important precisely because introduced pests can have significant impacts. The categorisation of pests is also highly political because it affects the level of government funding that is committed to managing the pest, making it critical that stakeholders are satisfied with the decision process. In a complex decision there are many issues that might be disputed.

The report demonstrates how formal group aggregation methods (both 'behavioural' and mathematical methods) can be employed to help settle distinct components of a decision problem. The types of issues that may need to be settled by a group are divided into a number of categories to assist the decision-maker in choosing the appropriate group aggregation method for a particular context. For instance, different situations warrant either a single person being responsible for an estimate or else a group being responsible; issues addressed by groups might concern either scientific facts or values; group members might be expected to submit honest opinions in some situations while in others they might be expected to 'play strategically'. The main role of the report is to provide guidance as to how to determine what group methods are appropriate in different circumstances.

The focus on situating formal 'consensus' methods within a broader decision problem and on matching methods to particular circumstances distinguishes this survey report from others of a similar kind. Previous surveys in the literature list the range of consensus methods that have been developed, and discuss their mathematical properties (where relevant), but they do not consider the methods in a genuine decision setting, or provide much detail about the more pragmatic properties of the methods.

1.3 Recommendations

Expand the existing survey:

- Consider in more detail the variety of methods that fall under the banner of "Bayesian updating". The methods listed in the report are drawn from one of the leading surveys of formal consensus methods (Clemen and Winkler 1999), but there are other methods that are commonly referred to as "Bayesian updating". It would be useful to consider the variety of Bayesian methods applied to one or two specific case studies.
- Do some explicit testing of the 'strategy proof' merits of the 'Lehrer-Wagner operationalised' method for assigning weights to group members for linear/logarithmic pooling.
- Consider the various measures of the accuracy of experts' probability judgments and critically evaluate Cooke's method for assigning weights to group members on the basis of this kind of information. Suggest alternative methods in the spirit of Cooke's approach.
- Consider how to track and represent uncertainty associated with group opinions.

Implement selected methods from survey:

- Develop the materials necessary for implementing some selected methods from the survey. This may require a computer program.

2 Introduction

2.1 The Broad Decision Context—A Case Study

We present a decision problem to illustrate the sort of contexts in which formal 'consensus' models can play a role. The case study used here is the decision framework outlined in ACERA Project 0707 Report entitled "Prioritising the impact of exotic pest threats using Bayes nets and MCDA methods". This section outlines the basics of the case study.

Risk assessments or prioritisation lists provide tools that can be used to support the exclusion of invasive species as well as to assess the potential impact of those that have become established. The objectives of such lists are usually to estimate a relative ranking of risk based on a prediction of whether or not a species is likely to be invasive and what the impact of an invasion would be. In the model developed in ACERA Project 0707, the probability of spread of a pest once established and the overall probability of spread include an assessment of the feasibility of control. These are examined using a Bayesian net approach. Multi-criteria decision analysis (MCDA) is then used to assess environmental, economic and other social impacts of pest spread. This impact assessment is subsequently combined with the probability of spread to give a score that can be used to prioritise pests according to their potential impact (the higher the score, the more serious the risk posed by the pest). The final score for pest *i* ($Score_i$) is calculated as follows:

$$Score_i = Pr(Spread_i) \times Impact_i$$

where
$Pr(Spread_i)$ is the overall probability of spread of pest *i* determined using a Bayes net and $Impact_i$ is the overall impact of pest *i* determined via MCDA.

Figure 1 presents the Bayes net that was developed to account for the factors that contribute to the probability of spread. These include the probability of, respectively, entry to Australia, establishment and spread upon arrival. Table 1 presents the basic multi-criteria decision table for assessing the impact of a pest. The relevant criteria to be considered are: economic cost (or benefit), conservation areas affected, indigenous and other protected areas affected, and public opinion. The type of multi-criteria method that was applied in the case study was a weighted linear average method. The overall impact, $Impact_i$, for each pest is just the weighted average of its impacts according to each of the 4 criteria. The formula is as follows:

$$Impact_i = \sum_{k=1}^{M} Z_k(Pest_i) \times c_k$$

where

$Z_k(Pest_i)$ is the normalised score of option $Pest_i$ under impact criterion *k* and c_k is the normalised weighting of importance of impact criterion *k*

Figure 1 Hypothetical Bayesian net analysis to estimate the probability of spread of an invasive pest showing criteria considered in assessing potential spread of a pest in Australia (Burgman M. et al., University of Melbourne, 2007). 'True' / 'False' and 'High' / 'Low' represent possible states against which a percentage probability of the state is specified, based on data or expert judgement.

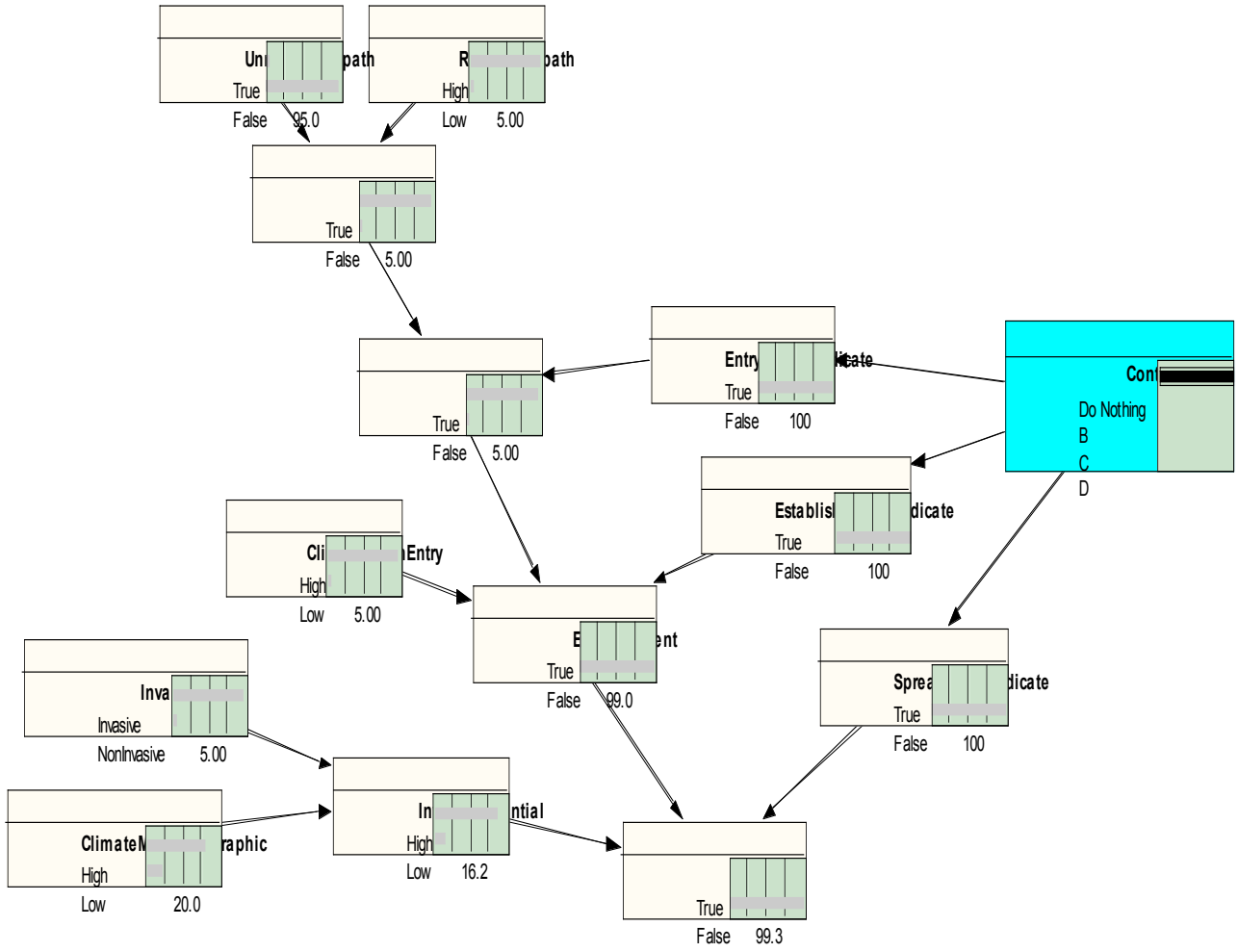


Table 1 **Multi-criteria Decision Table for assessing the impact of various plant pests in terms of 4 main criteria (with sub-criteria).**

Criteria	Alternatives				
	$Pest_1$	$Pest_2$	$Pest_3$	...	$Pest_n$
Environment					
Conservation area affected					
Economy					
Cost					
Social					
Indigenous area affected					
Public concern					

Note that, like any modelling exercise, there is a trade-off between simplicity and accuracy in this decision model. Simple models are often efficient to use. But they may involve some crude approximations and omit details that, if fixed, would improve the accuracy of the result. In this case, one might want a more detailed assessment of the potential impacts of a pest, e.g. economic cost might be broken down into a number of sub-criteria that include things like cost/gain to agriculture, health costs etc. Additionally, it would be better if the cost of managing a pest (and perhaps even different levels of management) could be separated out as opposed to being bound up in the overall probability of spread for a pest. These moves would add much complexity to the model, however.

Assume then that the decision agent has settled on the overall structure of the decision problem, that is, all relevant parties agree that the Bayesian net in Figure 1 is the appropriate way to determine the overall probability of spread, and the multi-criteria decision model in Table 1 is appropriate for determining the expected impact of a pest, should it succeed in spreading. It is useful to list the critical variables that appear in the Bayes net and the multi-criteria table. The following list also contains a brief description of how the variable should be evaluated (as indicated in ACERA Project 0707 Report). The probabilities of entry, establishment an spread depend on assessments of the feasibility of control.

Bayes Net input: Probability of Entry (PoE_i)
This is the probability that pest *i* enters Australia. It depends on the global distribution of the pest and its proximity to Australia, the potential entry sites, and the frequency of use of relevant international trade routes.

Bayes Net input: Establishment ($Pr(Est_i)$)
This is the probability that pest *i* successfully establishes in Australia. It depends on the climate match at the point of entry. Of course, a pest can only become established if it has gained entry to Australia.

Bayes Net input: Potential spread ($Pr(Spread_i)$)
This is the probability that the pest spreads from the point of establishment to its full potential range. A pest can only spread if it has become established. Spread also depends on the invasiveness potential of a pest, which in turn depends primarily on the climate match with the pest's native range and whether it is invasive elsewhere.

Multi-criteria input: Level of public concern (P_i)
The ACERA Project 0707 Report suggests that for each pest, the level of public concern can be estimated from the number of Australian media articles found on the internet over a 5 year period 2001–2006 that focus on the pest.

Multi-criteria input: Score for Economic Cost/Benefit (EC_i)
The information used to score each pest against the economic criteria can be sourced from published literature or industry or health cost statistics available on web pages (see ACERA Project 0803 Report). So as to avoid 'double dipping', i.e. double counting an impact in the economic criteria and again in the environmental or public

opinion criteria, it is recommended that the environmental, cultural or indirect social costs of a pest be excluded from economic considerations.

Multi-criteria input: Score for Conservation area affected (C_i)
Conservation area affected is estimated as the proportion of conservation area to total Australian landmass that would be affected by the uncontrolled spread of the pest.

> In order to calculate C_i, two variables must be determined. These are the spread distribution of a pest (DS_i) and the distribution of conservation areas (DC). C_i is then just the area overlap of DS_i and DC over the total Australian landmass.

Conservation area potentially affected does not evaluate the breakdown of different areas affected (e.g. Ramsar treaty wetlands; other significant wetlands; areas forming Regional Forest Agreements, National Parks, Crown land). Moreover, this criterion does not address the number of indigenous species that could decline as a result of the pest spreading into the Australian landscape.

Multi-criteria input: Indigenous or other protected areas affected (I_i)
This criterion is estimated as the proportion of indigenous (or other) area out of the total Australian landmass that would be affected by the uncontrolled spread of the pest.

> As per the criterion above, to calculate I_i, the decision maker needs to know the spread distribution of a pest (DS_i) and the distribution of indigenous areas (DI). I_i is then just the area overlap of DS_i and DI over the total Australian landmass.

The number of heritage listed or iconic sites that would be affected is not considered in the assessment.

Multi-criteria input: Weightings for criteria (c_k)
These are the weights of importance for each of the criteria in the multi-criteria decision table. As the name suggests, these weights should reflect the relative importance of the criteria (given the way in which options are scored against the criteria).

2.2 When Group Opinions Matter

With reference to the case study described above, a number of quantitative values must be settled to use the decision model for prioritising plant pests: PoE_i, P_i, C, etc. The basic question is: what is the appropriate way to settle these values? Those responsible for the decision (hereafter referred to collectively as the 'decision agent' or 'decision-maker') might be legally bound to follow some specific procedures (which could be critically examined in the light of the considerations raised in this report). Otherwise, the decision agent should simply aim for the best decision possible. It is also advisable that they aim for a 'stable' decision, or in other words, a decision that will not be challenged by disgruntled stakeholders. One can never guard

against this possibility completely, but if the decision model is constructed in a transparent way and choices are justified by data or by reference to other, similar decisions, the final result is likely to be more persuasive.

As far as justification goes, just setting up the basic decision model (as per Figure 1 and Table 1 above) is an important start. Making the different components of the decision explicit provides a focus for debate. It also makes apparent that the decision (in this case the ranking of pests) depends on a diverse range of factors—the decision agent must gather data from a variety of sources, and they will need to call upon different sorts of experts to settle different aspects of the decision model. The pest prioritisation problem involves biology/ecology-related questions about the potential spread of pests to estimates of potential economic gains/losses from any predicted pest spread. Any one person is unlikely to have the necessary expertise for settling all the issues that are relevant to the decision.

It is useful to distinguish between two different aspects of the case study decision model—those values that are part of the structure of the decision model and are effectively constants (hereafter referred to as decision *parameters*), and those values that are specific to the individual pests under assessment (referred to as decision *inputs*). To give a couple of examples, the weights of importance for the various criteria in the multi-criteria table are decision parameters, while the probability of entry for a pest is a decision input. The reason for making the distinction between the two sorts of values is to emphasise that only some issues—the decision inputs—should be evaluated on a pest-by-pest basis. The values of the decision parameters must therefore be carefully chosen because they will affect the outcome of every pest assessment. Indeed, the parameter/input distinction might mark a further broad division of labour in the decision process—those responsible for setting up the decision framework need not be responsible for supplying the inputs for particular cases. Such a division of duties might help to reduce bias in the pest prioritisation process; it might prevent the special interests of particular stakeholders from influencing the assessment of individual pests.

Another significant distinction vis-à-vis decision models is the fact/value distinction. It is reasonable to think that factual and value issues are best settled in different ways—indeed, Section 3 of this report treats the two types of issues separately. The most clear-cut value consideration in the Bayes net/multi-criteria model is the weightings of importance for the criteria (*C*). All other decision parameters/inputs are factual issues, or at least, they depend to a large extent on the facts. The scoring of pests' impacts on conservation and indigenous areas could be considered a mixed fact/value question. In any case, value judgments can (or perhaps always) enter into the interpretation and presentation of facts. Experts are socially situated, and their perception/communication of the facts will typically be affected by their own interests (see Burgman pp. 88–90). This motivational bias cannot be completely avoided (and it need not be malicious or conniving). But we can mitigate the influence of values on factual judgments. Indeed, this is one of the challenges for eliciting and combining expert opinions, and for decision modelling in general.

Return to the decision framework for prioritising pests that is depicted in Figure 1 and Table 1. There are various ways the decision-agent might go about settling the requisite decision parameters/inputs. An initial consideration is whether a single

person should have ultimate responsibility for the value of a parameter/input, or whether a suitably chosen group should have shared responsibility for the parameter/input. When it comes to determining a group opinion, the first case is much more straightforward, as it is not really a group exercise at all (even though it might involve the consultation of a number of experts). The second case raises many more problems because it is unclear just what a 'group opinion' really amounts to. First, however, it is useful to consider what sort of contexts would warrant each kind of approach.

A single person's opinion is reasonable when the responsible individual is recognised to be more or less impartial—at the least, the person must not stand to benefit personally from any particular result, and at best, the person would have a positive incentive to arrive at the objectively correct result. It is also important that the individual be an expert on the issue in question, or else able to do the requisite research, which may involve consulting others.

There may not be such an individual. Particularly if the issue is rather complex and 'interdisciplinary', it may need to be debated within a group setting. Moreover, some situations might call for more explicit demonstration of impartiality and democratic procedures, and in such cases it is advisable that the opinion of a group be sought.

The following table suggests a division of the decision parameters/inputs relevant to the case study into those that are best decided by a single responsible person (presumably within a single government agency), and those that are best handled by an independent committee. Two different types of groups are listed—the *expert* group and the *political* group—corresponding to the fact/value distinction. A brief justification for the three-way division of the parameters/inputs is given in the right-hand column. Note that decision parameters appear in *Arial* (blue) font and decision inputs appear in *Baskerville* (red) font in the table. The decision inputs have subscript *i* to indicate that these are values specific to some pest *i*.

Table 2

	Parameter/Input	**Justification**
Expert Single	*Level of Public Concern:* P_i *Probability of Entry:* PoE_i	One salient data source Government is impartial and has primary expertise, but benefit from consulting data and experts.
Expert Group	*Probability of Establishment:* $Pr(Est_i)$ *Probability of Spread:* $Pr(Spread_i)$ *Spread Distribution:* DS_i *Score for Economic Cost:* EC_i *Distribution of Conservation Areas: DC* *Distribution of Indigenous &*	Scientifically controversial. Better to consult outside expert group (which may include government members).

	Protected Areas: DI	
Political Group	*Criteria Weights: C*	Political issue requiring representation from different community groups.

For the three main categories of opinion in the table above, there are various methods for achieving a final result. The various methods are listed in the table below. In the next section of the report, these methods will be discussed in turn, with reference to the decision parameters/inputs from the case study.

Table 3

Individual/Group	**Details**	**Methods**	**Section**
Single Agent	Consults single data source	N/A	3.1
	Consults number of experts/data sources	• Bayesian updating	3.1
Group Agent (Expert)	Opinion of group of experts	• Behavioural methods • Mathematical methods	3.2
Group Agent (Political)	Opinion of group of political reps.	• Behavioural methods • Mathematical methods	3.3

3 Survey of Formal 'Consensus' Methods

3.1 Single Expert Gathering Evidence

This sub-section discusses situations in which a single person has responsibility for a factual parameter/input in a decision model. This may be appropriate if the person in question is suitably impartial and possesses the necessary expertise for seeking and subsequently assessing the relevant evidence upon which to base their final evaluation of the decision parameter/input. This is the least problematic case as far as group decision-making goes because it is not really a group decision at all.

3.1.1 Estimate based on single data source

The simplest case is where the responsible person bases their opinion on a single salient indicator of the decision parameter/input in question (this is the first entry in Table 3). In the case study, the score for one of the criteria in the multi-criteria model—*Level of Public Concern (P)*—might be best decided in this way. For instance, a plausible metric is the number of Australian media articles found on the internet over the 5 year period 2001–2006. Of course, this metric could be improved—for instance, one could check for duplicated articles reprinted by various media outlets, or else distinguish between articles supporting the establishment of the pest versus those opposing it. But the simpler metric may well be justified in terms of efficiency, and it is reasonable to assume that a single person could take responsibility for doing the counting of articles.

3.1.2 Evidence from a number of data sources

The single responsible person may make a final assessment of some decision parameter/input after considering evidence from a range of sources. (In this section we exclude the case where the sources are other experts; that is the topic of the next section.) A standard scenario is one in which an agent wants to estimate the characteristics of a population (for instance the population mean), given a number of samples. Each sample provides some information about the population as a whole. The question is: how should one combine the information from a number of samples in an appropriate way?

A situation of this sort might arise in relation to the case study being considered. The person responsible for determining the probability of entry into Australia for a particular pest i (PoE_i) might hold that a good initial estimate of this decision input is the probability of pests from the same region as pest i entering Australia in any given year (and assumed to be constant from year to year); call this probability p. The value of p will not be known exactly, but there may be a variety of sample data (number of pests detected by quarantine in a given year over the number of pests that were exposed to the relevant ports). Sample sizes may differ.

There are two large classes of statistical methods for making inferences about the characteristics of a population from sample data—frequentist/classical (e.g. Neyman-Pearson) methods and Bayesian methods. There are many textbooks devoted to statistical practices of both kinds. A number of books and articles compare the two approaches: see, for instance, Gigerenzer (1993), Howson and Urbach (1989) and Sober (2008, chap. 2). Here we will outline a simple application of Bayesian methods for combining various frequency data.

3.1.2.1 Summary of method

The Bayesian begins with a *prior* probability function over the hypothesis space. The hypothesis space might be continuous; for instance, a researcher might want to know the probability of extinction of a species within the next 10 years, with the hypothesis space being the interval [0, 1]. In this report, however, we will consider only discrete hypotheses. In this case the hypothesis space can be represented as a set of *m* hypotheses {*H1, H2,…*, *Hm*}. The agent's prior probability function is then:

$$\{Pr(H1), Pr(H2), \ldots, Pr(Hm)\}$$

The Bayesian method recommends that an agent should update their probabilities upon learning some evidence *E* in such a way that their subsequent or *posterior* probability $Pr_2(Hi)$ for each hypothesis *Hi* is equivalent to their prior probability for *Hi* conditional on *E*:

$$Pr_2(Hi) = Pr(Hi|E) = Pr(E|Hi).Pr(Hi) / Pr(E)$$

The evidence could be of any kind whatsoever (and it need not be in favour of the hypothesis—the posterior probability of *Hi* might be lower than its prior probability). Here we are interested in cases where the hypotheses concern some characteristic(s) of a population, and the evidence pertains to sample data.

3.1.2.2 Example

The decision maker wants to estimate the past probability of entry into Australia of a particular pest in a given month. A number of hypotheses are under consideration, and each are attributed equal prior probabilities, as follows:

Hi	*PoE = 0.2*	*PoE = 0.3*	*PoE = 0.4*	*PoE = 0.5*	*PoE = 0.6*
Pr(Hi)	0.2	0.2	0.2	0.2	0.2

The following four samples are recorded:

Year	**1**	**2**	**3**	**4**
# Pest Species exposed to port	30	40	28	10
# Pest Species entered Aust.	10	15	17	4

The Bayesian researcher could update their prior probabilities for the hypotheses with respect to each of the four samples sequentially, or they could just combine the samples into a mega-sample. The mega-sample says that the proportion of pests exposed to ports that entered Australia in a year is

(10 + 15 + 17 + 4) / (30 + 40 + 28 + 10) = 46 / 108

This sample proportion (46 out of 108) is the evidence *E*. It is used to update the probabilities for each of the hypotheses according to the Bayesian formula above. The posterior probabilities thus calculated can be found below:

Hi	***PoE = 0.2***	***PoE = 0.3***	***PoE = 0.4***	***PoE = 0.5***	***PoE = 0.6***
Pr(E|Hi)	0.0000001	0.0016877	0.0666453	0.0235718	0.0001015
Pr_2(Hi)	0.0000006	0.0183430	0.7243559	0.2561978	0.0011028

Note that for each *Hi*, the value of *Pr*(*E* | *Hi*) is calculated according to the binomial formula, e.g.

Hi = 0.3; *E* = 46 detected out of 108; $Pr(E \mid Hi) = {}^{108}C_{46} \times 0.3^{46} \times 0.7^{(108-46)}$

$= 0.0016877$

We can see from the results that [*PoE* (per annum) = 0.4] is the most likely hypothesis ($\times 0.7$).

Note that the final estimate for the probability of entry for the pest will depend on the duration of the period in question. The above probabilities are for a single year, so for *n* years the probability must be multiplied by *n*. (The accuracy of the overall probability of entry for a particular period of time would be improved if the test probabilities were attached to smaller time intervals e.g. *PoE/month* or *PoE/day*. We ignore that complication here.) Note that after a sufficient number of years it is practically certain that the pest will gain entry into Australia.

3.1.2.3 Discussion

It must again be acknowledged that Bayesian methods are being given prominence here, as opposed to frequentist methods for drawing inferences from sample data. Some claim that Bayesian methods are flawed precisely because of the place they give to a decision maker's prior beliefs about the probabilities of the various hypotheses under consideration—this is regarded as subjective data. Bayesians emphasise that subjective opinions influence any kind of risk assessment, and that it is in fact a virtue of the Bayesian model that it makes this aspect of the reasoning process transparent. Note that for the example above, the *likelihoods*—the values of *Pr*(*E* | *Hi*)—are objective, as they are calculated according to the binomial formula.

3.1.3 Evidence from a number of experts

Bayesian methods can also be used to update one's opinion about a decision parameter/input in response to the subjective opinions of other experts. More

precisely, the Bayesian model, to be described in detail below, stipulates how a rational agent should update their *prior* beliefs upon learning the opinions of a number of experts. When it comes to standard risk analysis, Clemen and Winkler (1999, p. 190) report that French (1985), Lindley (1985) and Genest and Zidek (1986) "all conclude that for the typical risk analysis situation, in which a group of experts must provide information for a decision maker, a Bayesian updating scheme is the most appropriate method".

The Bayesian model for updating in response to expert opinions can be applied to any kind of parameter/decision input. The decision maker might want to estimate a point value, a probability distribution, or indeed they might be entertaining any kind of hypothesis about some state of affairs in the world. The evidence submitted by the experts can also take a variety of forms. For instance, assume the decision maker wants to estimate the population size (k) for some species. Experts might contribute their best point estimate for k, or else a probability distribution over the possible values for k, or some other data that has a bearing on the value of k. Typically, however, the experts are asked to submit opinions of the same form as the decision-maker. For example, in this case, if the decision-maker wants a point-estimate of population size (k), then the experts are asked to submit point-estimates of k. The Bayesian method stipulates how the decision-maker should update their own estimate for k in light of the experts' opinions.

3.1.3.1 *Summary of method*

The decision maker has some prior probability distribution over a set of hypotheses. (The "prior" probability of the decision maker may already incorporate frequency data from a variety of sources, as per the methods of the previous section.) This is denoted by $Pr(H)$, where $H = \{H1, H2, \ldots\}$. There are n experts. According to the Bayesian model, the decision-maker should update $Pr(H)$ in response to the opinions of the experts (where the combined opinions of the experts are expressed as D) to their prior conditional probability $Pr(H|D)$. As mentioned above, it is generally assumed that the experts provide opinions that have the same form as the decision-maker's. For instance, if the decision maker is interested in the probability of extinction of a species, then the experts submit their estimates of this very parameter—the probability of extinction of the species—as opposed to some other relevant information like the minimum viable population of the species. In what follows we will assume that the decision maker and the experts are estimating $Pr(H)$. This is to say that D, the summary of the experts' opinions, has the following form (where $Pr_i(H)$ is expert i's subjective probability distribution over the hypothesis space):

$$D = \{Pr_1(H), Pr_2(H), \ldots Pr_n(H)\}$$

We can apply the Bayesian formula (which exploits Bayes' rule) to get the decision maker's final or posterior probability distribution over the hypothesis space, given they have learnt D:

$$Pr(H|D) = Pr(D|H).Pr(H) / Pr(D)$$

The formula itself is quite straightforward. The problem is that it may be very difficult to specify the terms in the right-hand expression that are needed for computing $Pr(H|D)$. Clemen and Winkler (1999, pp. 191–194) describe a number of off-the-shelf models for the relevant probabilities, some suited to the scenario in which experts submit a single event probability, and others suited to situations in which experts submit a probability distribution over a continuum of hypotheses.

3.1.3.2 Example

Bayesian updating might be employed to refine the probability of entry (*PoE*) of a pest into Australia. Assume that the individual interviews a number of experts who each gives him/her a probability of entry for the pest in question. The individual may have already taken into account frequency data from a number of sources, as per the methods of the previous section. Now the hypotheses are {Entry, No Entry}, however, as opposed to $\{PoE = 0.2, \ldots, PoE = 0.6\}$.

Let PoE_i ($i = 1, \ldots, n$) denote expert i's stated probability that the pest enters Australia. Given the reports of the experts, the decision-maker might update their initial probability of entry for the pest, PoE_0, via one of the models surveyed by Clemen and Winkler (1999) that can be expressed as follows:

$$\frac{PoE_f}{1 - PoE_f} = \frac{PoE_0}{1 - PoE_0} \prod_{i=1}^{n} \frac{\Pr(PoE_i \mid entry)}{\Pr(PoE_i \mid \sim entry)}$$

Note that the final result is expressed as an odds ratio. Note that $Pr(PoE_i \mid \text{entry})$ is the probability that the decision-maker gives to expert i estimating that the probability of entry is PoE_i given that the pest does in fact enter Australia. Likewise, $Pr(PoE_i \mid \text{no entry})$ is the probability that the decision-maker attributes to expert i submitting a probability of entry of PoE_i given that the pest does not enter Australia. These conditional probabilities thus indicate the decision-maker's opinion of the accuracy of expert i.

The model just described is appropriate for scenarios in which the experts each bring independent information to the problem of assessing a pest's probability of entry (*PoE*). For example, if all experts say that the probability is 0.6, then the decision-maker's final probability, PoE_f, will tend to be much higher than 0.6 (depending on the decision-makers prior probability, PoE_0). Whether or not this is appropriate depends on the sorts of experts that are interviewed. But we assume here that the experts do in fact have different background evidence—one expert might be a quarantine official, another an authority on agricultural trade, another might be a scientist who is informed about the current world distribution of potential pests, and so on. The decision-maker's own prior probability, PoE_0, might just be the proportion of pests that were investigated in the past that turned out to have entered Australia.

The following table shows the data that is necessary for determining a final probability of entry for a pest using the above model, for a situation in which 4 experts are interviewed. (Example numerical values for the decision inputs are provided. Note that the prior probability for the pest entering Australia is given as 0.4. This in fact corresponds to the hypothesis with the greatest posterior probability in the previous section.)

Table 4

Expert Data	**Decision-maker's probabilities**	
	$PoE_0 = 0.4$	
$PoE_1 = 0.65$	$Pr(PoE_1 \mid \text{entry}) = 0.7$	$Pr(PoE_1 \mid \text{no entry}) = 0.3$
$PoE_2 = 0.85$	$Pr(PoE_2 \mid \text{entry}) = 0.8$	$Pr(PoE_2 \mid \text{no entry}) = 0.5$
$PoE_3 = 0.50$	$Pr(PoE_3 \mid \text{entry}) = 0.2$	$Pr(PoE_3 \mid \text{no entry}) = 0.5$
$PoE_4 = 0.70$	$Pr(PoE_4 \mid \text{entry}) = 0.5$	$Pr(PoE_4 \mid \text{no entry}) = 0.4$

Using the values given in the table, the decision maker's final odds ratio for the entry of the pest, $PoE_f / (1 - PoE_f)$, can be calculated according to the above model as follows:

$$PoE_f / (1 - PoE_f) = 0.4/0.6 \times 0.7/0.3 \times 0.8/0.5 \times 0.2/0.5 \times 0.5/0.4 = 1.24$$

$$PoE_f = 0.55$$

Note that the probabilities given by the experts are not included in the expression that gives the final probability of entry. The values that do enter into this expression are the probabilities that the decision-maker attributes to the experts giving the values that they do, conditional on the pest entering and not entering Australia.

3.1.3.3 Discussion of Bayesian updating

The features of the Bayesian model that have attracted praise when it comes to updating in response to expert opinions are the very same features that are criticised by others. Many consider it a major strength of the Bayesian model that it is supported by a well-worked-out rationale (as compared to the averaging models that are discussed in the next section), and that it can handle interdependencies or correlations between experts' probability judgments. Note that the model above assumes that expert opinions are independent, i.e.

$$Pr(Pr_1(H), Pr_2(H), \ldots Pr_n(H) \mid H) = Pr(Pr_1(H) \mid H) \times Pr(Pr_2(H) \mid H) \times \ldots \times Pr(Pr_2(H) \mid H)$$

The joint conditional probability need not be equivalent to the product of the individual conditional probabilities, however. Correlations of any kind between expert opinions can be represented in the Bayesian model.

The problem with the Bayesian model for updating on expert opinions is that even in the simplest case where expert judgments are assumed to be independent, as per the example above, the Bayesian model is rather difficult to use. In particular, the likelihood of a particular expert's opinion given that the hypothesis is true ($Pr(PoE_i \mid$ entry) for the case above) and the likelihood of that same expert's opinion given that the hypothesis is false ($Pr(PoE_i \mid$ no entry) above), are very difficult to interpret and evaluate. (This is in contrast to the likelihoods employed in the Bayesian updating of the previous section, which were determined by the binomial formula.) Yet these

expert opinion likelihoods play a crucial role in the calculation of the final/posterior probability for the hypothesis. One could say that the generality/flexibility of the Bayesian model outlined above comes at a cost—the decision maker must make some rather tricky assessments of the opinions expressed by other experts. Nonetheless, these kinds of Bayesian models for updating opinions in response to the opinions of a number of experts are regularly employed in risk assessment applications (see, e.g., Clemen 1985, Clemen and Murphy 1986, Clemen and Winkler 1987, Roosen and Hennessy 2001).

There may be alternative sorts of models suitable for combining expert opinions that are Bayesian in spirit. For instance, there may be models that treat individual expert opinions as sample data, in line with the kind of Bayesian updating that is discussed in 3.12. Models of this sort are not discussed in the current key surveys on combining expert opinions, such as Clemen and Winkler (1999). The possibility of alternative Bayesian methods for combining expert opinions is something that should be investigated further.

3.2 Opinion of an Expert Group

Sometimes the most appropriate way to settle an issue is to seek the opinion of an expert group. There might be reason to think that the group is more likely to 'get the facts right', especially when the relative competence of the individual experts is unknown. Or else, it might be important for political reasons that a decision, or an important parameter/input to a decision model, be the opinion of a group rather than a single individual. This would make the decision seem more democratic and impartial, and thus less likely to be later challenged by disgruntled interest groups.

This section considers methods for forming a group opinion, or in other words, methods for aggregating individual opinions to get an overall group opinion. The decision parameters/inputs from the case study that are most appropriate for illustrating these methods are ones that either fall outside the decision-maker's direct expertise, or else are politically sensitive issues for which it is wiser to have group involvement. The first sort of methods considered here are structured processes for group deliberation known as *behavioural methods*. Some of the more complex decision parameters/inputs in the case study may be best decided in this way: the 'set-up' for scoring *Conservation Areas Affected (C)* and *Indigenous Areas Affected (I)*, and, on a pest-by-pest basis, the predicted distribution of a pest that is used to determine its actual scores for *C* and *I*. The second kind of methods considered here are more rigid procedures that exploit mathematical algorithms; they can be referred to as *mathematical methods*. The decision inputs *Establishment (E)* and *Spread (S)* are used to illustrate this latter group of methods.

Behavioural and mathematical group aggregation methods are often depicted as being in competition with each other. For instance, Clemen and Winkler (1999) and O'Hagan et al. (2006) compare the performance of these two types of methods in terms of whether the group arrives at the right result (how well they 'track the truth'). It seems more apt, however, to regard the two method-types as complimentary. Behavioural methods recommend strategies of communication within a group. In other words, behavioural methods consider the psychology of group members, and

techniques for controlling their interaction so as to get the best decision results. Most behavioural methods depend on a facilitator of some kind, and might be interpreted as methods for group facilitation, but the end result is supposed to be a group decision. The facilitator simply guides the group through the process of sharing data and the reasoning behind individual estimates. At the end of such a process, however, the opinions of group members may still differ, and at this stage it might be useful to employ a mathematical method to achieve a common group opinion. By the same token, there is no reason to think that mathematical methods for reaching a group decision necessarily preclude prior discussion and the opportunity for individuals to learn from others.

A distinction that is important to both behavioural and mathematical methods is that between a group *consensus* and a group *compromise*. A consensus is a group opinion that is shared by all members of the group. In other words, a consensus occurs when all members of the group either have, or come to have, the same opinion on some matter. A compromise, on the other hand, could be described as a situation in which group members 'agree to disagree'; the group members differ in their individual opinions, but they settle on some group opinion as being properly representative of the group (see Steele et al. 2007 for more on the distinction). Note that even in the latter case, there must be a consensus about something—a compromise can only be reached if members share the same view about what is the appropriate method for determining the group opinion. While methods for aggregating opinions to achieve a group opinion are often referred to as *formal consensus methods*, this label is somewhat misleading. No method can guarantee a group consensus (in the absence of significant assumptions about group members' attitudes about each other's opinions). There is always the possibility that group members will disagree at the end of the day, whether out of stubbornness or for entirely legitimate reasons. Moreover, Peterson et al. (2005) argue that just assuming that group discussion will lead to consensus is a dangerous ideal that will generally lead to the wishes of the dominant few being forced upon the rest. Some presentations of behavioural methods do not appreciate this fact. But the presumption of consensus need not be a feature of any method, as will become evident in what follows.

3.2.1 Behavioural methods

Behavioural methods garner support from psychological studies showing the problems that arise in unstructured group discussion. Such problems are: A dominant group member can manipulate group members to reach a position these other members do not hold (Hamilton 2003, Steinel and De Dreu 2004); the formation of social cliques within the group can isolate and alienate other group members that have unique expertise (Thomas-Hunt et al. 2003); and idiosyncrasies of group size and group member status can lead to deference to a single group member irrespective of that member's depth of knowledge (Ohtsubo and Masuchi 2004). There are further studies from Stasser and Titus (1985) and Wittenbaum and Stasser (1996) showing that unsupervised groups can be poor at identifying and pooling specialist information held by individuals. Others have found that group opinions can become polarised around extreme values, depending on the dynamics of the discussion. Studies from Janis (1982), Plous (1993), Sniezek (1992) and Heath and Gonzalez (1995) confirm this phenomenon, referred to as 'group overconfidence'.

Of course, there is nothing wrong with group members updating their opinions on the basis of others, and this can result in the group having a more extreme opinion than that of any individual. For example, it may be perfectly legitimate for the group to assign a probability of 0.9 to *X* if all members assign a probability of 0.7; perhaps the members are perceived to have independent background information about *X* that, taken together, makes *X* very probable. The evidence suggests, however, that in many cases unstructured groups are pushed to extreme opinions for entirely non-epistemic reasons.

Behavioural methods are intended to alleviate the biases that arise in unstructured group discussion. Innami (1994) finds that "the quality of group decisions increases to the extent that group members exchange facts and reasons and decreases to the extent that group members stick to their positions, and that an intervention that emphasizes a knowledge-based logical discussion and consensual resolution of conflicts improves the quality of group decisions" (reported by Clemen and Winkler 1999, p. 197). It is no small task to design a group process that yields a "knowledge-based logical discussion". As mentioned, aiming for *consensual* resolution of conflicts may be counterproductive to this goal—we shouldn't force agreement where there is none to be had. Putting the consensus ideal aside, however, the idea is to assess behavioural methods in terms of how well they lead to group decisions based on reasoning rather than bullying in particular contexts.

Behavioural methods are very versatile—they can be used to arrive at a range of different kinds of group outputs, not just probability distributions or point estimates. Indeed, there are some rather complex decision parameters/inputs from the case study that may be best settled via the use of behavioural methods. For the sake of this example, we take the two decision parameters that provide the framework for scoring a pest's impact with respect to *Conservation Areas Affected (C)* and *Indigenous Areas Affected (I)*. We are assuming that *C* is effectively the proportion of conservation area to total Australian landmass that would be affected by the uncontrolled spread of the pest. In order to calculate *C* for individual pests, the decision-makers must determine the distribution of conservation areas in Australia (call this *DC*). This is a one-off decision, which is why it can be referred to as a decision *parameter*. Likewise, we will assume that *I* is the proportion of indigenous area to total Australian landmass that would be affected by the uncontrolled spread of the pest. In order to calculate *I*, the distribution of indigenous lands in Australia (call this *DI*) must be settled. Again, this is a decision parameter—it is constant for all pests under assessment. To calculate $C(\text{pest}_i)$ and $I(\text{pest}_i)$, we need to know the expected distribution of the pest: $DS(\text{pest}_i)$. The group outputs in each case (*CI*, *DI* and $DS(\text{pest}_i)$) will be distribution maps. Before considering the examples themselves, however, it is useful to describe the main behavioural methods in general terms.

3.2.1.1 Summary of methods

Several behavioural methods will be briefly outlined here: to begin with, a brief description of market-based methods will be given. *Prediction markets* involve very detached groups in which members do not communicate directly with one another yet nonetheless update their opinions in response to the opinions of others. The remaining methods all involve group members being brought together in a more formal setting

to share the reasons upon which their opinions are based. Call these the "discursive behavioural methods". Three methods of this sort that will be considered are: the Delphi approach, the Nominal Group Technique and the Closure method.

Prediction markets. Prediction markets are speculative markets that are used for making predictions. Participants place bets on whether or not an event will occur, or whether a parameter takes a particular value. The current market prices are then interpreted as the probability of the event, or they can be used to calculate the expected value of the parameter in question. Prominent defenders of these markets include Surowiecki (2004) and Sunstein (2006). There is evidence to suggest that prediction markets are accurate predictors of events/parameter values, but some are sceptical about these results and about just what market prices represent vis-à-vis participants' beliefs (see, for instance, Manski 2006). More investigation is necessary to determine precisely what conditions are likely to yield predictions with a specified accuracy (including betting conditions and incentives, as well as the competence and independence of experts on the issue in question). Surowiecki (2004, p. 10) suggests that crowds will be "wiser" to the extent that there is diversity of opinion and independence amongst bettors, among other things.

Delphi approach. Note that the discursive behavioural methods which follow are in many ways very similar; all presuppose a forum for group discussion, and begin with a process of problem formulation and the provision of background information and context to experts. Some authors provide very explicit instructions regarding this initial setting-up process (see, e.g., Vose 1996). The main differences amongst the discursive methods have to do with the level of anonymity that is maintained after the initial set-up stage vis-à-vis the opinions and arguments submitted by group members.

According to the Delphi method, each individual anonymously submits their opinion regarding some unknown value/parameter (whether a probability distribution, a single-point estimate, or something else) together with a brief explanation of the opinion. The group might also be supplied with statistics describing the overall distribution of opinions: for instance, the median and the interquartile range. Group members do not interact in any other way. Individuals can update their own estimate based on this information about the opinions of others. This updating process is iterated until individuals no longer wish to revise their own estimate. Some presentations of the Delphi method assume that it ends in consensus, but this need not be the case.

Nominal Group Technique. This method is similar to the Delphi method except that there is allowance for group discussion at each iteration after individuals have submitted their new opinion.

Closure method. This method (developed by Valverde 2001) stipulates that group discussion should focus on the relationship between experts' opinions, rather than on the reasons for each expert's opinion. Initially, each expert advances a number of claims, which may be rebutted by other experts (whether the target is the data, the model or the broader background theory underlying the claims). Experts must then formulate their positions in more precise terms. The subsequent group discussion should focus on locating the reasons for disagreement amongst experts—there could be some ambiguity about what it is that is being evaluated, or there could be

disagreement about how to interpret the data, or about what model and background theory should be appealed to. The idea is that pinpointing sources of disagreement makes it more likely that experts will resolve their differences and come to consensus. The experts may, however, simply agree to disagree.

3.2.1.2 Examples

Consider first *CS* and *IS*. The location of conservation and indigenous lands across Australia is, in one sense, not something that requires group involvement. The decision-maker need only consult the appropriate public records and cadastral databases to construct distribution maps of these two types of special-status lands. On the other hand, when it comes to assessing the impact of pest species, a more inclusive definition of "conservation" and "indigenous" lands might be more appropriate. For instance, perhaps conservation areas are not just designated national parks and state reserves, but also other public lands and private property that have high conservation value. In such case, and also when it comes to classifying indigenous lands, an expert group might be assembled to determine a suitable classification system and distribution map.

There is both a political and a scientific/historical component to assessing what lands are important for conservation/indigenous reasons. This is to say that the relevant experts are identified by their political *and* scientific credentials. For instance, whether or not some land area should be classed as *indigenous land* is presumably to a large extent a matter of whether the relevant political group—indigenous Australians—regard the land as culturally important. There might also be a historical requirement—a need for evidence of the sustained importance of the land in question. Similar sorts of political-scientific issues will arise in the identification of *conservation land*. Indeed, many public decisions will be of this mixed fact-value nature, and in such cases the public acceptability of the decision will be largely determined by the choice of expert group; there needs to be representation from the appropriate stakeholders and knowledge groups. Beyond the mere formation of an appropriate expert group to decide upon such issues, it would be desirable, of course, if the group functioned well.

In cases like this where group members are handpicked in order to achieve the appropriate political representation, market-based methods for arriving at a group conclusion are inappropriate. Prediction markets are suited to cases in which a large number of people have an incentive to bet on an issue, and where relevant information is scattered throughout the group. Moreover, the precise workings of prediction markets are not well understood, and there would not be sufficient justification for deciding upon quasi-political issues in this way. The point of the group being comprised of representatives from different social sectors is to have these representatives share with each other the views and interests of their respective social groups. So some form of group discussion is required. Given the evidence cited above regarding the problems with unstructured group discussion, there is reason to employ one of the discursive behavioural methods.

It is clear from the summary of the discursive behavioural methods above that the emphasis is on an *iterated* process of opinion updating, and the major distinction between methods is the level of anonymity. When it comes to highly politicised issues

like identifying conservation and indigenous lands, arguably anonymity is not very useful. The experts in the group are presumably selected because they represent different knowledge/cultural groups, so it would probably be quite obvious who is the author of the various submitted opinions and arguments. In such case, it would not only be very hard to keep the group process anonymous, but it would also seem rather dishonest. The group is more likely to be satisfied with the decision process if members have the opportunity for face-to-face debate about their respective positions. In other words, the *Nominal Group Technique* or the *Closure Method* would be more appropriate than *Delphi*. (The *Closure Method* is perhaps more developed than the *Nominal Group Technique* in that it recommends the group focus on dispute resolution.) The iterated process is useful because it prevents rambling discussion and keeps all group members involved—at regular intervals, members have the opportunity to consider all the arguments on their own terms, and make an individual assessment of the state of play which is then communicated to the group. The ideal situation is when group members come to consensus on an issue, or at least negotiate a compromise in the final iterations.

There is also the spread distribution for each pest, $DS(\text{pest}_i)$, to decide upon. This is more clearly a straight scientific question, so the expert group would be selected on the basis of knowledge of the pest in question and broader plant ecology. (A market-based approach might also be useful for answering this question, provided there were enough participants who had incentives to bet or who were willing to make hypothetical bets.) If an expert group were to be used, the optimal group would contain a range of experts whose knowledge/skills compliment one another. It is not clear what would be the best way to conduct the group discussion—if there looked to be overbearing personalities within the group or persons of high scientific standing that others would tend to unquestioningly defer to, then the anonymous *Delphi* method could prove useful; if the group seemed naturally very participatory, then the *Nominal Group Technique/Closure Method* would probably work well.

3.2.1.3 Discussion

It is important to emphasise that there is a lot more to say on the issues raised in this section. Market-based group processes are still not well documented or properly understood and further research in this area would be desirable.

When it comes to the discursive behavioural methods, the psychology of group interaction is complex, and has been the topic of much research. The idea here is simply to introduce the main discursive methods and their basic principles for effective group discussion. It is difficult to establish which method will be most successful in a particular group situation. Experimental conclusions about the value of anonymity are mixed. For instance, Myers and Lamm (1975) report evidence that face-to-face interaction in groups working on probability judgments may lead to social pressures that are unrelated to group members' knowledge and abilities. On the other hand, depriving the group of open discussion may stifle the sharing of information and lead to inferior group judgments, especially when it comes to more complex issues.

3.2.2 Mathematical methods

Mathematical methods for aggregating opinions are in a sense much more restrictive than behavioural methods. They propose a specific algorithm for combining opinions, rather than allowing group members to arrive at consensus or compromise via any old path of opinion-change and negotiation. This might be considered an advantage; mathematical methods are reliable in the sense that particular group member inputs lead to a particular group output. So if individual members input the same opinions in two different decision scenarios, the group output will be the same. Some view this feature of mathematical models as a disadvantage, however—the idea is that the rigidity of mathematical methods undermines effective group reasoning because it tries to force an algorithm onto a process that is inherently non-mechanistic.

It is likely that there is no formula for constructive group discussion. Having said that, there is no reason to think that mathematical methods for reaching a group decision necessarily preclude prior discussion and the opportunity for individuals to learn from others. Moreover, a behavioural method might be selected to facilitate this initial process. As noted above, however, it is dangerous to expect that a behavioural method will end in consensus, i.e. all group members having identical opinions on the issue at hand. At the point where group members are reluctant to change their own opinions and yet there still exist differences of opinion within the group, a rigid mathematical method is arguably the best way to achieve group *compromise*.

While group discussion prior to determining a compromise position is the ideal, sometimes there will be reason to sidestep discussion and employ mathematical methods from the outset. The previous section discussed a range of behavioural methods that accommodate differing amounts of face-to-face contact. In some situations, it might be thought most beneficial to have no dialogue at all within the group. Perhaps the opportunity for dialogue would lead most group members to defer to a few senior experts within the group, due to lack of confidence. In such cases, the best way to reach an opinion that is representative of the group might be to have group members immediately submit their individual opinions, which are then combined according to the chosen mathematical algorithm.

3.2.2.1 Summary of methods

The main mathematical methods for aggregating group member opinions are averaging methods. There are two dominant weighted average methods:

- weighted linear average, or “linear pooling”
- weighted logarithmic average, or “logarithmic pooling”

Weighted averages can be applied to different types of numerical inputs, including point estimates. A problem that often arises in group contexts is the *allocation problem*. This is the problem of deciding how a fixed amount of some currency (whether it be money or probability or something else) should be divided/distributed among a set of alternatives. More formally, an allocation problem amounts to determining the values in an array, where these values must add to some positive real number S, and all values in the array must themselves be positive real numbers. We will refer to such arrays as *allocation arrays* or *allocation distributions* (Wagner 1982). There are numerous examples of allocation problems, including:

- allocating weights of importance to criteria (e.g. in a multi-criteria decision problem—refer to the next Section)
- distributing a fixed amount of money to various projects
- distributing probability over a given state space

The mathematical details of the two dominant weighted average methods (where the inputs/outputs are allocation arrays) are as follows:

Linear: $$a(k) = \sum_{i=1}^{n} w_i a_i(k)$$

Logarithmic: $$a(k) = r \prod_{i=1}^{n} a_i(k)^{w_i}$$

Where:
$a(k)$ is the k^{th} element of group array a
$a_i(k)$ is the k^{th} element of individual i's array a
w_i is the weight attributed to person i (where the weights for all n members add to 1)
r is a normalising constant

We might want to compare the merits of linear and logarithmic averages when it comes to aggregating probability distributions. (Arguably the most common kind of quantitative value that we want a group to decide upon is a probability distribution.) There has been some investigation of this issue in the literature. The properties that are considered desirable are listed in the table below.[1] The latter three can be described as unanimity properties—if a method has such a property, it is to say that when everyone in the group is in agreement about something, then the resultant group opinion is also in agreement. For example, we might be interested in preserving unanimity regarding the probability of an event (*unanimity*) or, in a multivariate setting, unanimity regarding the independence of events (*independence preservation*) or unanimity regarding mutual exclusiveness of events (*coherent marginalisation*). The first criterion—*externally Bayesian*—concerns whether timing matters with respect to forming a group opinion; it would be preferable if the final group opinion was the same whether or not the group formed before or after some new piece of evidence became known to all group members.

Table 5

Property	**Linear**	**Logarithmic**
Externally Bayesian*: when new data is obtained, updating the previously pooled group distribution equates to updating the individual group members' distributions and then pooling these.*	no	yes
Independence preservation*: If all group members find that two propositions are independent such that Pr(A and B) = Pr(A).Pr(B),*	no	yes

[1] The results in the table are discussed in Genest and Zidek (1986), Clemen and Winkler (1999) and O'Hagan *et al.* (2006).

then the pooled group distribution should also find the two propositions to be independent.		
***Coherent marginalisation**: If all group members find that two propositions are mutually exclusive such that Pr(A or B) = Pr(A) + Pr(B), then the group should also find the two propositions to be independent.*	yes	no
***Unanimity*: If all group members agree on the probability of some event, then the group probability of the event should equal this common value of members.**	yes	no (but satisfies zero unanimity)

It is interesting to analyse the properties of linear and logarithmic averaging, but unfortunately the summary given in the table above does not suggest one or the other pooling method to be superior. Both methods have just two of the stated properties. In any case, the relevance of these properties for comparing opinion pooling methods has been questioned by French (1985), Lindley (1985) and Genest and Zidek (1986) on the grounds that a group should not be expected to behave like a single agent. A more basic point is just that it is unlikely that the typical group decision scenario will involve a complex probability function for which judgments of independence or mutual exclusiveness amongst events is important.

The critical issue when it comes to using either of the two averaging methods is the choice of weights for group members. The table below lists some suggestions in the literature for assigning weights, which will be discussed later in relation to specific examples. The choice of weights does not change what properties (amongst those in Table 5) an averaging method has, except for the case where one group member receives the maximal weighting of one and the others receive a weighting of zero. In this case, the group opinion just is the opinion of the chosen individual, both for linear and logarithmic pooling. Funnily enough, this is the best group aggregation method by the lights of the properties listed in Table 5 because it has all four of the listed properties. But of course this does not seem to be a genuine group aggregation method (and indeed it is typically stipulated that all group members receive at least some positive weight so that their opinion makes at least some difference to the group's opinion).

Table 6

Weighting Method	**Comment**
Cooke's performance	Well recognised difference in expertise, perhaps based on past performance in similar decision scenarios.
Best expert takes all	Might be achieved through a vote, or recommended by past performance data
Equal weights	Might be recommended by past performance data. Alternatively, there may be no basis for differences in

	expertise.
Lehrer-Wagner	Arguably achieves "consensus"
Lehrer-Wagner operationalised	May avoid strategic play

When it comes to factual issues, arguably the choice of weightings for a group aggregation method should be entirely based on achieving the most accurate group result (the group result that is most likely to 'track the truth'). The problem is that it is generally very difficult to make this assessment. Each of the weighting distributions listed in the table above has its merits.

Cooke's weightings. Any method that assigns weights relative to performance might be referred to as a version of Cooke's method. For instance, there might be public records of experts' past performance (according to some measure of accuracy) on similar factual issues that could govern the current distribution of weights. The original version of Cooke's method involves a test to elicit experts' competence with respect to the issue in question (see Cooke 1991). Experts are asked to evaluate variables for which the true value of a number of instances of the variable is known (but unknown to the expert). For each variable, the expert indicates the probability that it falls within a certain region; for instance, the expert might indicate the 5th, 50th and 95th percentiles for the variable. Experts are weighted according to their relative performance, where this is based on the calibration and information content of their probability assignments. The weight for expert j is proportional to the product of a calibration component C_j and an information component K_j. Both components are based on the idea of a Kullback-Leibler (K-L) distance between two discrete probability distributions. Let $\mathbf{p} = \{p_1, p_2, ..., p_m\}$ and $\mathbf{q} = \{q_1, q_2, ..., q_m\}$ be two probability distributions over a state space of size m. Then the K-L distance between them is:

$$I(p,q) = \sum_{i=1}^{m} p_i \ln(p_i/q_i)$$

The calibration component C_j is based on $I(p_j, q_j)$, where q_j indicates expert j's probabilities for regions of the variable space and p_j is the proportion of true values of the variable that fall in each of the probability regions elicited from expert j. The information component K_j is defined as the K-L distance between the expert's distribution, q_j, and a uniform distribution. A more informative distribution will be far from uniform, placing concentrations of probability on relatively short ranges.

Best expert takes all. While Cooke's method of assigning weights relative to experts' past performance makes intuitive sense, we might question whether this is the best way to utilise the past performance data. If there are records showing which expert is the most accurate, why not just base the group opinion on this expert's opinion? After all, if the aim is to achieve the most accurate group result possible, it is not clear that a differentially weighted average of member opinions will be more accurate than the best expert's opinion. Winkler and Poses (1993) show that it is, in fact, very difficult to make any general claims about what will be the most accurate combination of group member opinions—not only does it depend on the measure of accuracy that is used to test past performance, but also the correlations between group member opinions are important.

Equal weights. As per the previous paragraph, performance data might indicate that an equally weighted group would have been more accurate than any individual on their own. Again, this is to say that past performance data need not recommend that weights be assigned relative to individual performance. The data might suggest that an equally weighted group would outperform the sort of arrangement recommended by Cooke. There is another quite different argument for equal weightings: in the absence of any public data regarding the performance of group members, it is most natural to assign experts equal weights.

Lehrer-Wagner weights. Lehrer and Wagner (1981) propose a method that allows weights to be decided within the group rather than imposed on the group. This could be particularly useful in cases where past performance data is either non-existent or ambiguous, and where the assignment of equal weights does not seem satisfactory. Initially, each member assigns their own set of weights to all group members; this data can be summarised in a matrix M where row i corresponds to individual i's distribution of weightings. Provided M satisfies certain conditions,[2] there will be some n such that M^n is a matrix with equivalent rows:

$$M^n = \begin{bmatrix} w_1 & w_2 & \dots & w_n \\ w_1 & w_2 & \dots & w_n \\ \dots & \dots & \dots & w_n \\ w_1 & w_2 & \dots & w_n \end{bmatrix}$$

These rows represent the group distribution of weightings. Lehrer and Wagner tell a more elaborate story as to why it is appropriate to determine the group distribution of weightings in this way; their story models the group process as one of individual updating to reach *consensus* as opposed to mere *compromise*. Some have disputed the generality of the model as a model of consensus; indeed, one would not expect every expert group to come to consensus. The model might also be interpreted, however, as a method for the group to determine the weighting distribution for a compromise.

Lehrer-Wagner operationalised. Regan et al. (2006) proposed a modification of the Lehrer-Wagner method for determining the distribution of weights across experts. According to this method, weights are derived from the experts' estimates of the numerical value of interest: group member i's weight for group member j (w_{ij}) is assigned based on the distance j's view (p_j) is from i's view (p_i). The suggested formula for calculating this weight is as follows:

$$w_{ij} = \frac{1 - |p_i - p_j|}{\sum_{j=1}^{n} 1 - |p_i - p_j|}$$

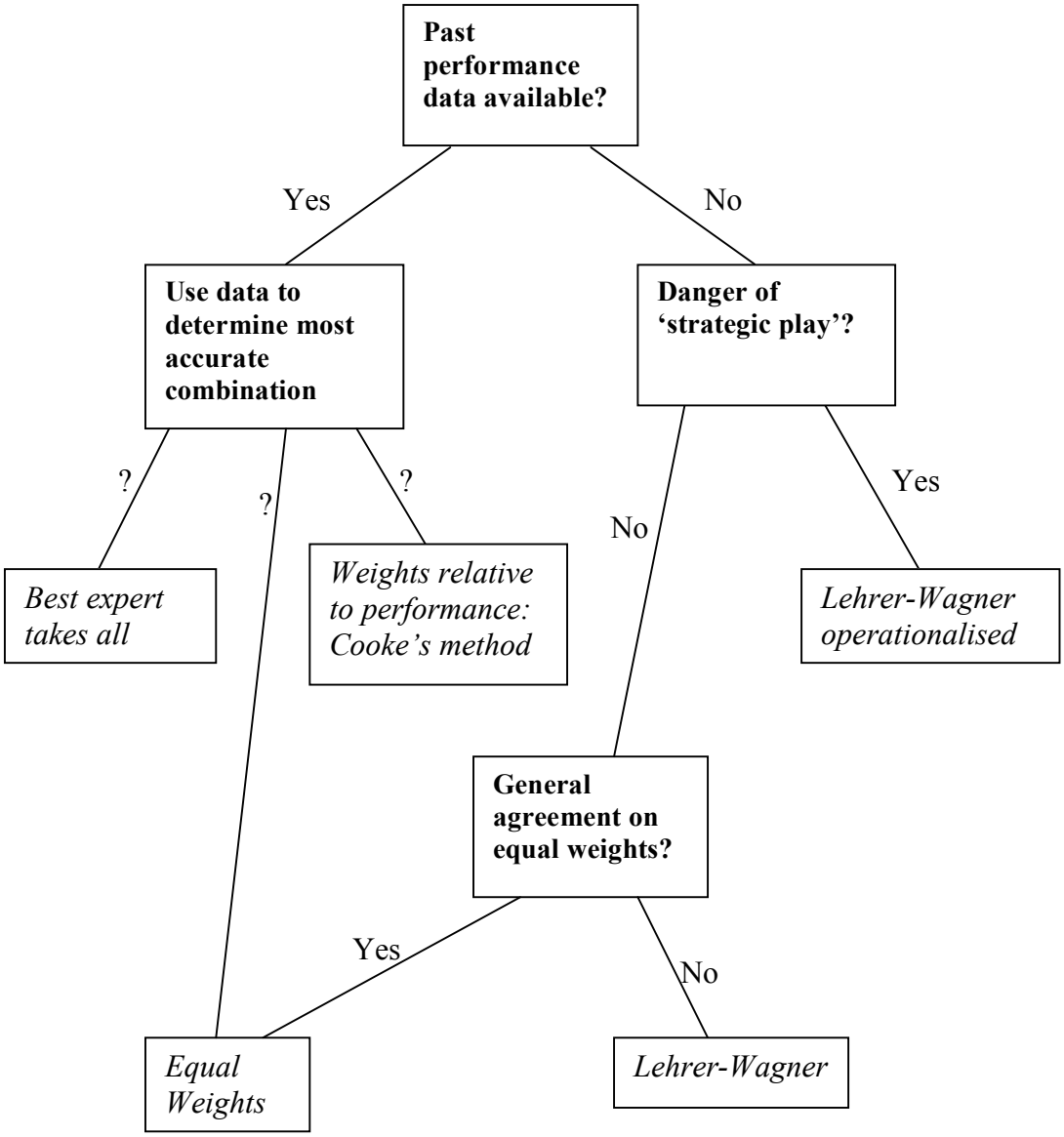
[2] Each individual has to be connected to all others via a 'chain of respect' and someone has to assign him/herself positive weight. Note that person i is connected to person j via a 'chain of respect' if i assigns positive weight to someone who assigns positive weight to someone else and so-on down the chain until someone assigns positive weight to j.

Note that the above formula has the effect of down-weighting 'extreme' opinions—those opinions that deviate largely from the opinions of most group members. Once the original matrix of weightings is determined according to the above formula, the method proceeds as per the regular Lehrer-Wagner method. There is some n such that M^n has rows that are equivalent and these are taken to be the group's weighting distribution. Like the regular Lehrer-Wagner method, the weights are thus determined from within the group—there is no assumption of prior consensus regarding the appropriate distribution of weights. The chief benefit of the 'operationalised' approach is that it allows the computation of initial weightings based solely on the individuals' probability estimates. This may be distinctly advantageous in practical situations where group members are unwilling or unable to quantify their respect for the competence of others in the group—it is one thing to acknowledge a relative ordering of respect for other individuals in a group, it is an entirely different matter to place an abstract numerical value on the levels of respect a person has for the views or expertise of other members of the group.

Another advantage of this method is that it makes 'strategic voting' very difficult. The other methods for assigning weights are independent of the experts' opinions on the factual issue of interest. This means that experts might have the incentive to state their opinion dishonestly, if they predict that this will help to compensate for the other group members' opinions shifting the group result too far in the 'wrong' direction. When weights depend on the distances between expert opinions, the further i's opinion is from j's, the less weight they give each other, so it is unclear how either member should go about playing strategically.

The following flow chart is intended to assist a group in determining the appropriate choice of weights for use in linear/logarithmic pooling.

Figure 2



To illustrate the rationale for choosing a weighting scheme, as depicted in the flowchart in Figure 3, it is helpful to refer to a couple of examples from the case study.

3.2.2.2 Example

Two issues that were tagged earlier as being best decided by an expert group are the probability that are particular pest ($pest_i$) will become established in Australia ($Pr(Est_i)$), and the probability that it will spread uncontrollably ($Pr(Spread_i)$). These two issues are related (note that the probability of spread given that the pest has not become established is zero):

$$Pr(Spread_i) = Pr(Spread_i \mid Est_i) \cdot Pr(Est_i)$$

We can also expand the last term to take account of the probability of entry in Australia, such that the entire expression is as follows (again, the probability of establishment if the pest does not enter Australia is zero):

$$Pr(\text{Spread}_i) = Pr(\text{Spread}_i \mid \text{Est}_i) \times Pr(\text{Est}_i \mid \text{Entry}_i) \times Pr(\text{Entry}_i)$$

Assume that the probability of the pest entering Australia has already been determined. (This was covered earlier—it served to illustrate how a single person may negotiate, via Bayesian updating, a range of different evidence bearing on some issue.) What remains is to determine $Pr(\text{Spread}_i \mid \text{Est}_i)$ and $Pr(\text{Est}_i \mid \text{Entry}_i)$ for the pest in question. These are quite distinct from the issue of whether a pest will enter Australia. The latter has to do with trade routes and quarantine vigilance, while establishment and spread depend more directly on facts about plant biology/ecology. It is thus reasonable to think that the relevant probabilities should be determined by different individuals/groups.

It is plausible that the one group of experts will be appropriate for deciding both $Pr(\text{Spread}_i \mid \text{Est}_i)$ and $Pr(\text{Est}_i \mid \text{Entry}_i)$. These are arguably best decided by a group, as compared to an individual, because they are complex scientific issues that have a significant affect on policy, and so it is better that they be the shared responsibility of a group of suitable experts, for both democratic and accuracy-related reasons. In the initial problem formulation stage of the process, it should be noted that:

$$Pr(\text{Spread}_i \mid \text{Est}_i) \times Pr(\text{Est}_i \mid \text{Entry}_i) = Pr(\text{Spread}_i \mid \text{Entry}_i)$$

The group must finally decide upon $Pr(\text{Spread}_i \mid \text{Entry}_i)$, but it is presumably useful to think about this as a product two components—the components being the terms on the left-hand-side of the expression. On the other hand, there may be a tendency for experts to mistake $Pr(\text{Spread}_i \mid \text{Est}_i)$ for $Pr(\text{Spread}_i \mid \text{Entry}_i)$, which is to say that they may prefer to estimate $Pr(\text{Spread}_i \mid \text{Entry}_i)$ directly, rather than separate it into a product of two components. Perhaps the best way forward is to distinguish the two terms of the product during initial group discussion, and then concentrate exclusively on the final spread probability when it comes to aggregating the group member opinions to achieve a compromise.

It is reasonable to think that face-to-face group discussion would be helpful in this case. This is to say that the *Nominal Group Technique/Closure Method* would be more suitable than *Delphi* when it comes to facilitating the initial process of sharing data and arguments. Face-to-face discussion allows more detailed analysis of an issue, and this is important for negotiating complex issues that involve experts with a range of expertise and background knowledge. It also allows the opportunity to get to the bottom of experts' reasoning and the source of disagreements. In the spirit of the *Closure Method*, the group might find that they disagree about one of the terms in the product, say $Pr(\text{Spread}_i \mid \text{Est}_i)$, but not the other. The group could then focus their discussion on the disputed value.

After structured group discussion, disagreement might persist regarding the value of the product—$Pr(\text{Spread}_i \mid \text{Entry}_i)$. At this point it would be appropriate to employ a mathematical method to achieve a group compromise. Either linear or logarithmic pooling would be justified. The important question, in either case, is the choice of

weights for the group members. The group would do well to follow the chain of reasoning given in Figure 3. The first question is whether there are performance data available. In this case, it is likely that there is no possibility of finding or creating meaningful performance data. Even if the same experts had been involved in the assessment of a number of pests, arguably the individual cases are not sufficiently similar to warrant general conclusions regarding an expert's competence at assessing $Pr(\text{Spread}_i \mid \text{Entry}_i)$. In other words, each pest presents a unique problem due to its particular attributes and ecological niche, and an expert who has proven competent at assessing, say, *Rubus fructosis* (blackberry), may just as well be quite incompetent when it comes to assessing the potential spread of *Acacia nilotica* (prickly acacia).

In the absence of past performance data, the next question is whether experts can be relied upon to submit their honest opinions, or whether they are likely to act strategically so that the final group result is more likely to resemble their true opinion about $Pr(\text{Spread}_i \mid \text{Entry}_i)$ for a particular pest (pest_i), or else more likely to support their preferred final outcome. Given that the prioritisation of plant pests is a highly political issue that affects industry and has significant social/economic consequences, there is certainly the possibility that members of the expert group will have an interest in the policy consequences of their collective opinion, and so will try to influence the group result. The operationalised Lehrer-Wagner method might thus be appropriate for determining the weights to be used in linear/logarithmic pooling.

For example, assume that there is a group composed of 6 experts who are each asked to submit the probability of spread for, say, *Rubus fructosis* (blackberry), given that it has entered Australia. The left-hand vector represents the true opinions of the experts. The right-hand vector represents the opinions that the experts actually submit.

	0.19		0.19
	0.09		0.09
⇨	0.53	⇨	0.83
	0.15		0.15
	0.22		0.22
	0.11		0.11

Notice that the third expert submits a different probability from their actual estimate of the probability of spread for *Rubus fructosis*. This might be because expert #3 is very concerned about the effect on biodiversity of this potential weed, and they think that the other group members grossly underestimate its probability of spread, so they try to compensate for this by submitting a much higher estimate than they would otherwise submit if all group members shared their opinion that the probability of spread 0.5.

The operationalised Lehrer-Wagner method determines weights for the group members in such a way that expert #3's submitted opinion does not have the compensating effect that they might have intended it to have. The reason for this is that expert #3 receives less weight than other group members (and so has less influence on the final group probability) because the distance from expert #3's opinion to anyone else's opinion is very large.

Before looking at the operationalised Lehrer-Wagner calculations, it is useful to first consider what the group probability for spread would be if equal weights were used in the weighted linear average. Consider first what the result would be if the experts all submitted their true opinion:

$$Pr(\text{Spread}_i \mid \text{Entry}_i) = \begin{matrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{matrix} \begin{matrix} 0.19 \\ 0.09 \\ 0.53 \\ 0.15 \\ 0.22 \\ 0.11 \end{matrix} = 0.215$$

Now consider what the group result would be if equal weights are used and expert #3 submits a much higher probability than what they think is the actual probability:

$$Pr(\text{Spread}_i \mid \text{Entry}_i) = \begin{matrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{matrix} \begin{matrix} 0.19 \\ 0.09 \\ 0.83 \\ 0.15 \\ 0.22 \\ 0.11 \end{matrix} = 0.265$$

Expert #3 would be happier with the group probability being 0.265 as opposed to 0.215 (because the former is much closer to their true estimate of 0.53), so there is incentive for this expert to 'play strategically' if equal weights are assigned to all group members.

It was decided, however, that the operationalised Lehrer-Wagner method for assigning weights is most appropriate. The weighting matrix, calculated via the distance via formula given above, is as follows:

$$M = \begin{matrix} 0.196 & 0.176 & 0.070 & 0.188 & 0.190 & 0.180 \\ 0.182 & 0.202 & 0.053 & 0.190 & 0.176 & 0.198 \\ 0.138 & 0.100 & 0.383 & 0.123 & 0.149 & 0.107 \\ 0.188 & 0.184 & 0.063 & 0.196 & 0.182 & 0.188 \\ 0.192 & 0.172 & 0.077 & 0.184 & 0.198 & 0.176 \\ 0.183 & 0.195 & 0.056 & 0.191 & 0.177 & 0.199 \end{matrix}$$

Note that all members (apart from expert #3) assign expert #3 very low weight. The matrix M^n that has equivalent rows (representing the 'group' distribution of weights) is as follows:

$$M^n = \begin{matrix} 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \\ 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \\ 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \\ 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \\ 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \\ 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \end{matrix}$$

The weighted linear average representing the group opinion with the above weighting array is:

$$Pr(\text{Spread}_i \mid \text{Entry}_i) = \begin{matrix} 0.183 & 0.178 & 0.093 & 0.184 & 0.181 & 0.181 \end{matrix} \begin{matrix} 0.19 \\ 0.09 \\ 0.83 \\ 0.15 \\ 0.22 \\ 0.11 \end{matrix}$$

$$=0.215$$

In this example, the group opinion that is calculated using operationalised Lehrer-Wagner weights and the probabilities that experts actually submit is equivalent to the group opinion that *would have* resulted from applying equal weights to the *true* probability estimates of the experts. Of course, the analysis need not have worked out like that—the true probability estimates of the experts could have been anything, and the Lehrer-Wagner result does not necessarily equate to the group opinion that would have resulted from true expert probabilities and equal weights. What is evident, however, is that it is very difficult to 'play strategically' to achieve one's desired group opinion when the operationalised Lehrer-Wagner method is used to determine weights. This particular example shows that grossly overstating or understating one's probability estimate does not have the desired effect on the group opinion due to the penalty in weightings that is imposed when one's own opinion is very distant from the others in the group.

3.2.2.3 Discussion

As stated, there is no entirely principled way to determine how weights (for achieving a group average) should be distributed across experts. It is advisable that the group conduct a sensitivity analysis of their choice of weights. For instance, in the example above, the probability of spread was calculated first using equal weights and secondly using the weights recommended by the operationalised Lehrer-Wagner method. In this case, the difference in the weighting arrays and the final group results (probability of spread) was not large, but possibly significant. The Lehrer-Wagner method was deemed preferable here, given that expert #3 was 'playing strategically'. It is apparent that the Lehrer-Wagner method is resistant to at least some kinds of 'strategic play', but more detailed analysis of the method should be conducted to determine what exactly are its merits in this respect.

3.3 Opinion of a Political Group

While there is often good reason to make factual issues the business of a group, when it comes to prioritising community values, it is more or less essential that any opinions relevant to public decisions are those of a suitably chosen group. Moreover, the choice of group is a delicate issue—members should represent the different interests of the entire population, yet they should also be willing to consider the views of others and think about what is best for the community at large.

The obvious value issue in the case study example is the choice of weights-of-importance for the various criteria in the multi-criteria model. (These can be denoted c_i to emphasise that they are distinct from the weights that are assigned to group members in the linear/logarithmic pooling algorithms.)The distribution of criteria weights is a *parameter* of the decision model—it is not expected to change on a pest-by-pest basis. Regardless of whether it is *Mimosa pigra* (mimosa) or *Lantana camara* (lantana), say, that must be assessed, the relative importance of the different criteria for scoring impact (ecological impact, economic impact, etc.) stays constant. This means that the assessment of individual pests is not a value-laden issue (although private interests can affect the assessments of spread probabilities and so on given by experts). Values only explicitly enter into the choice of criteria weights of importance, which are part of the general multi-criteria framework by which all pests are scored and subsequently ranked.

3.3.1 Methods

The kinds of methods that can be used by a political group to decide upon a value issue are the same as those that may be employed by an expert group deciding upon a factual issue. Both the behavioural and the mathematical methods of the last section serve as useful tools. The considerations of a political group in using these methods, however, will be somewhat different from those of an expert group. It is best to proceed straight to the example from the case study to illustrate.

3.3.1.1 Example

Given that the criteria weights are determined once only for the multi-criteria decision model, there is special incentive here to aim for the optimal group process. In this setting, the behavioural methods play a particularly important role because the final group compromise will be more stable the closer the group members' opinions are to each other. In other words, the more the members' come to understand each other in the discussion process, the more broadly acceptable the final compromise.

As mentioned, it is important that all the relevant segments of the community are represented in the group. In this case it is reasonable to include one/some conservationists, indigenous persons, industry representatives and farmers. For the sake of simplicity, assume that the group is compromised of one representative from each of these stakeholder groups. Given that group members will be likely to guess who belongs to each opinion/argument even if this information was kept anonymous, a behavioural method that allows face-to-face discussion seems most reasonable. The *Closure Method*, with its focus on locating individual differences, seems most

appropriate. There is likely to be a mix of factual and value disputes in deciding the relative importance of the criteria (even though we have classed this as a 'value' issue), and it would be extremely desirable for the group to work out whether they disagree on matters of fact or value or both.

It is worth noting for this particular example that an important part of the initial discussion of problem formulation and context is to ensure that group members are aware of how the various criteria will be scored. Steele et al. (to appear) emphasise that the criteria weights of importance within a multi-criteria model are meaningless if they are assessed independently of the scoring scales for the criteria. This is made clear by the fact that a change in scoring scale (for instance, let all the scores for economic cost be multiplied by 1/5) will lead to a change in the overall score for an option, and may lead to rank reversals amongst options. In the case study at hand, it is important for group members to know that *C* is assessed in terms of the proportion of conservation lands to total Australian land-mass that would be affected by the pest, and *EC* is inversely proportional to the estimated dollar cost (to give a plausible suggestion), where zero cost gets a score of one, and some proposed maximum cost (e.g. 5 million dollars) gets a score of zero. The scoring scales for the remaining two criteria are also important when it comes to assigning weights of importance.

The following table represents potential judgments about criteria weights (where these judgments are made in light of the scoring scales for the criteria). The initial judgments are unbracketed. Note that the values reported in the table are fictitious and are used here for illustrative purposes only.

Table 7: Group member assignments of criteria weights-of-importance

Group member	**Criteria**			
	Economic Cost (EC)	*Affect on Conservation Areas (C)*	*Affect on Indigenous Areas (I)*	*Public Concern (P)*
Conservationist	0.1 (0.2)	0.6 (0.5)	0.2	0.1
Indigenous rep.	0.2 (0.3)	0.3	0.4 (0.3)	0.1
Farmer	0.4	0.4 (0.3)	0.1 (0.2)	0.1
Industry rep.	0.6 (0.5)	0.2 (0.3)	0.1	0.1

There are significant differences between the initial weighting distributions that are proposed in Table 7. This is likely to be a common scenario—group members representing interest groups are likely to give a relatively large amount of weight to the criterion corresponding to their special interest.

According to the *Closure Method* (or a nearby variant of this method), the group members present their distributions of weightings together with reasons for their choice of distribution, to which other group members offer rebuttal. The aim of the iterated group discussion is to consider different possible sources of disagreement, and to work out whether these disputes can be resolved by clarifying terms, by persuasive argument, or whether they rather amount to genuine differences of opinion. It might be noted in this case that the group agrees at the outset that *Public*

Concern be given a weighting of 0.1, so this criterion can be set aside and discussion can be focussed on the remaining 3 criteria.

The discussion might lead group members to revise some of their criteria weights. Possible changes are represented by the bracketed values in Table 7 above. At this point, group members' opinions are stable (and yet they are not in consensus), so it is advisable that a mathematical method be used to reach a group compromise. Either linear or logarithmic averaging will do; as per aggregating expert judgments, the critical issue is the assignment of weights to group members.

If the group is composed of the right balance of community representatives, then it is reasonable to think that the assignment of equal weightings to group members is most appropriate. Fairness has particular importance when it comes to deciding matters of value; as French (1981, p. 332) comments, "in matters of preference, all men are equal; in matters of knowledge some are more expert than others". Referring to the flow chart in Figure 3: there is no sense in past performance data when it comes to value judgements, and in this case, it is reasonable to think that there is no opportunity for group members to play strategically, because their opinions are made public during the course of the discussion process. The assignment of equal weights to all members would thus be the most natural choice. Applying these weights to the linear pooling method gives the following group results for criteria weights-of-importance:

Economic Cost (EC):

$$0.25\square0.2 + 0.25\square0.3 + 0.25\square0.4 + 0.25\square0.5 = 0.35$$

Affect on Conservation Areas (C):

$$0.25\square0.5 + 0.25\square0.3 + 0.25\square0.3 + 0.25\square0.3 = 0.35$$

Affect on Indigenous Areas (I):

$$0.25\square0.2 + 0.25\square0.3 + 0.25\square0.2 + 0.25\square0.1 = 0.2$$

Public Concern (P):

$$0.25\square0.1 + 0.25\square0.1 + 0.25\square0.1 + 0.25\square0.1 = 0.1$$

So for the example outlined here, the compromise group position is that economic cost and affect on conservation areas be weighted equally, followed by affect on indigenous areas, followed by the level of public concern. As mentioned, to appreciate the significance of these criteria weights, one needs to know the scoring scales for options with respect to each of the criteria.

3.4 Conclusions

It is an important feature of this report that it depicts formal 'consensus' methods in their natural milieu, i.e. as part of a broader decision model. The case study presented here is relevant to biosecurity management and concerns the prioritising of non-indigenous non-primary industry pest threats, but the lessons apply more broadly. It should be clear that the formal decision modelling approach, as a package, has huge benefits—it introduces order to a complex decision problem and decomposes it into the variety of issues, both factual and value-based, that have a bearing on the final result. Each of these issues raises questions regarding the appropriate way to take into

account expert knowledge and resolve group disagreement. This report offers guidelines for choosing the 'right' formal consensus model in these various situations.

4 References

Baker, J. and Stuckey, M. (2008). "Prioritising the impact of exotic pest threats using Bayes nets and MCDA methods." ACERA Project 0707 Report.

Burgman, M. A. (2005). Risks and Decisions for Conservation and Environmental Management. Cambridge, Cambridge University Press.

Clemen, R. T. (1985). "Extraneous expert information." Journal of Forecasting **4**: 329–348.

Clemen, R. T. and A. H. Murphy (1986). "Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships." Weather and Forecasting **1**: 56–65.

Clemen, R. T. and R. L. Winkler (1987). Calibrating and combining precipitation probability forecasts. Probability and Bayesian Statistics. R. Viertl. New York, Plenum**:** 97–110.

Clemen, R. T. and R. L. Winkler (1999). "Combining Probability Distributions From Experts in Risk Analysis." Risk Analysis **19**(2): 187–203.

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

French, S. (1981). "Consensus of opinion." European Journal of Operational Research **7**: 332–340.

French, S. (1985). Group consensus probability distributions: A critical survey. Bayesian Statistics. J. M. Bernado, M. H. DeGroot, D. V. Lindley and A. F. M. Smith. Amsterdam, North-Holland. **2:** 183–197.

Genest, C. and J. V. Zidek (1986). "Combining Probability Distributions: A Critique and an Annotated Bibliography." Statistical Science **1**(1): 114–135.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in Statistical Reasoning. A handbook for data analysis in the behavioural sciences: Methodological issues. G. Keron and C. Lewis. Hillsdale, N. J., Lawrence Erlbaum**:** 311–339.

Hamilton, R. W. (2003). "Why do people suggest what they do not want? Using context effects to influence others' choices." Journal of Consumer Research **29**: 492–506.

Heath, C. and R. Gonzalez (1995). "Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making." Organizational Behavior and Human Decision Processes **61**: 305–326.

Howson, C. and P. Urbach (1989). Scientific Reasoning: The Bayesian Approach. La Salle, III., Open Court.

Innami, I. (1994). "The quality of group decisions, group verbal behavior, and intervention." Organizational Behavior and Human Decision Processes **60**: 409–430.

Janis, I. L. (1982). Groupthink: Psychological Studies of Policy Decisions and Fiascoes. Boston, Houghton Mifflin.

Lehrer, K. and C. Wagner (1981). Rational Consensus in Science and Society. Dordrecht, D. Reidel Publishing Company.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. Bayesian Statistics. J. M. Bernado, M. H. DeGroot, D. V. Lindley and A. F. M. Smith. Amsterdam, North-Holland. **2:** 375–390.

Manski, C. F. (2006). "Interpreting the Predictions of Prediction Markets." Economic Letters **91**(3): 425–429.

Myers, D. G. and H. Lamm (1975). "The polarizing effect of group discussion." American Scientist **63**: 297–303.

O'Hagan, A., C. E. Buck, et al. (2006). Uncertain Judgements: Eliciting Experts' Probabilities. Chichester, UK, John Wiley & Sons.

Ohtsubo, Y. and A. Masuchi (2004). "Effects of status difference and group size in group decision making." Group Processes and Intergroup Relations **7**(2): 161–172.

Peterson, M. N., M. J. Peterson, et al. (2005). "Conservation and the Myth of Consensus." Conservation Biology **19**(3): 762–767.

Plous, S. (1993). The psychology of Decision Making. New York, McGraw Hill.

Regan, H. M., M. Colyvan, et al. (2006). "A Formal Model for Consensus and Negotiation in Environmental Management." Journal of Environmental Management **80**(2): 167–76.

Roosen, J. and D. A. Hennessy (2001). "Capturing experts' uncertainty in welfare analysis: An application to organophosphate use Regulation in U.S. apple production." American Journal of Agricultural Economics **83**: 166–182.

Sniezek, J. A. (1992). "Groups under uncertainty: An examination of confidence in group decision making." Organizational Behavior and Human Decision Processes **52**: 124–155.

Sober, E. (2008). Evidence and Evolution. Cambridge, Cambridge University Press.

Stasser, G. and W. Titus (1985). "Effects of information load and percentage of common information on the dissemination of unique information during group discussion." Journal of Personality and Social Psychology **53**: 81–93.

Steele, K., Y. Carmel, et al. (to appear). "Uses and Misuses of Multi-criteria Decision Analysis (MCDA) in Environmental Decision-Making."

Steele, K., H. M. Regan, et al. (2007). "Right Decisions or Happy Decision Makers?" Social Epistemology **21**(4): 349–368.

Steinel, W. and C. K. W. De Dreu (2004). "Social Motives and strategic misrepresentation in social decision making." Journal of Personality and Social Psychology **86**(3): 419–434.

Sunstein, C. R. (2006). Infotopia: how many minds produce knowledge. Oxford, Oxford University Press.

Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many are Smarter Than the Few. New York, Doubleday.

Thomas-Hunt, N. C., T. Y. Ogden, et al. (2003). "Who's really sharing? Effects of social and expert status on knowledge exchange within groups." Management Science **49**(4): 464–477.

Valverde, L. J. (2001). Expert judgment resolution in technically-intensive policy disputes. Assessment and Management of Environmental Risks. I. Linkov and J. Palma-Oliveira. Dordrecht, Kluwer**:** 221–238.

Vose, D. (1996). Quantitative Risk Analysis: a Guide to Monte Carlo Simulation Modelling. Chichester, Wiley.

Wagner, C. (1982). "Allocation, Lehrer Models, and the Consensus of Probabilities." Theory and Decision **14**: 207–220.

Winkler, R. L. and R. M. Poses (1993). "Evaluating and combining physicians' probabilities of survival in an intensive care unit." Management Science **39**: 1526–1543.

Wittenbaum, G. M. and G. Stasser (1996). Management of information in small groups. What's Social about Social Cognition? Social Cognition Research in Small Groups. J. L. Nye and A. M. Brower. Thousand Oaks, CA, Sage.