Scientific expertise, risk assessment, and majority voting

Thomas Boyer-Kassem*

Working paper: comments welcome, but please do not quote without permission.

February 29, 2016

Abstract

Scientists are often asked to advise political institutions on pressing risk-related questions, like climate change or the authorization of medical drugs. Given that deliberation will often not eliminate all disagreements between scientists, how should their risk assessments be aggregated? I argue that this problem is distinct from two familiar and well-studied problems in the literature: judgment aggregation and probability aggregation. I introduce a novel decision-theoretic model where risk assessments are compared with acceptability thresholds. Majority voting is then defended by means of robustness considerations.

Keywords: scientific expertise, risk, majority voting, robustness, decision theory

*TiLPS, Tilburg University, The Netherlands. Email: t.c.e.boyer-kassem@uvt.nl

1 Introduction

Scientists are often asked by political institutions to give expert advice on pressing questions. For instance, agencies that regulate medicines regularly resort to expert panels, and national scientific academies give advice to the government or to the assemblies. Even after discussing, scientific experts do not always agree on the answer, and when they do, they may disagree on the justification for this answer. How should decisions that involve risk assessments be taken and justified within scientific expert panels? This is the central question studied in this paper. As a matter of fact, many expert panels take decisions using the majority voting rule. This is for instance the case in advisory committees in the European and in the American agencies that grant medicines authorization, respectively the EMA and the FDA.[1] But is it the best decision rule? Is majority voting *on the final decision* the best way to aggregate different experts' opinions, and to track their reasons? This paper is restricted to cases in which the expert panel is asked to take a decision on only one binary question, for instance to answer the question "Is the risk-benefit ratio of some medicine worth it to be authorized for commercial use?". This simple case is already interesting as it corresponds to many real-life cases: some expert panels are constituted on the sole purpose of answering one specific question, or are asked to answer several but logically unrelated questions — e.g. decisions about different medicines.

To study this problem, I introduce a novel decision-theoretic model. The true/false decision is supposed to be taken by comparing a risk assessment a (typically, a probability) to a risk acceptability threshold t, e.g. "true" if and only if $a < t$. For simplicity, a and t are supposed to be in $[0, 1]$, but any quantity might go.[2] It is assumed that the n experts agree on the threshold value, but differ in their individual risk assessments a_k ($k = 1, \dots, n$) — or conversely, that they agree on the assessment, but disagree on the threshold value. Typically, the question asked to the expert panel is in the form "Is X's risk below t?". The problem studied in this paper is to determine how the individual a_k's should be aggregated in comparison with t, so as to give the group's answer to this question (I shall speak equivalently of the group's decision, or of the group's belief on whether the risk is below t). Compared to probability aggregation theory which studies the aggregation of probabilistic opinions, the novelty of this model lies (i) in the introduction of a threshold comparison which projects probabilities into a binary space, and (ii) in the fact that the group has to take a

[1]Cf. Hauray and Urfalino (2007), Urfalino and Costa (2015).

[2]Real quantities can be mapped to the interval $[0, 1]$, for instance with the function $x \to 1 - 1/(1 + x)$).

stand on one binary question only, and not on a more complex agenda. Compared to judgment aggregation theory which studies the aggregation of an interconnected set of beliefs, the novelty is that individuals do not just have true/false beliefs but probabilistic ones, even if the group is asked to express a true/false belief in the end. The present problem can be considered as a first bridge between these two existing frameworks. The best decision rule for our binary question is likely to depend on the details of characteristics of the question, of the experts, of the available knowledge, and on other details. My methodological approach is not to conduct a detailed study of particular cases, but to look at features which most (interesting) cases share, so as to find general properties of the best decision rule — what is meant by "best" shall be discussed too.

The main claims of this paper are the following. I argue that the framework of probability aggregation cannot help us solve the present problem (Section 2), because the aggregation problems it considers are too general. For the aggregation of scientific risk assessment on a specific question, a theory of its own is needed, and I try to sketch one here. I then argue that robustness considerations clearly legitimate majority voting on the final decision (Section 3). But when justifications for the decisions are sought, majority voting can lead to inconsistencies and the expert panel should aggregate on the reasons separately, before deriving logically its decision (Section 4). Overall, the case for the majority rule is thus a mixed one.

2 Probability Aggregation and Beyond

A standard requirement for a scientific expert panel is that it provides justifications for its decision. In the present model, the decision has to be consistent with the comparison between the risk assessment and the threshold, so a minimal justification is that the panel has a belief on the risk assessment (as all experts have a belief on the risk assessment, it would be weird that the panel claims to refuse the authorization while not being able to say that it believes that the risk assessment is above the threshold). So, our problem includes as a first step the aggregation of the individual risk assessments $\{a_k\}_{1\leq k\leq n}$ into a single group assessment a — deeper justifications for the group's decision are contemplated in Section 4. The group's decision is supposed to be consistent with this assessment, so pragmatically the easiest way to do so may be for the group to first aggregate the individual assessments, and then compare the result to the threshold.

Majority voting on the decision itself is a standard way for expert groups to take decisions, but it does not proceed in that way. Can it be objected that, within our model, it lacks the requirement that the group should be attributed a belief on the risk assessment? No, for the following reason. The result of the majority vote is "true" if and only if a majority of agents vote "true", i.e. if and only if a majority of agents have a numerical assessment below the threshold, i.e. if and only if the median of the agents' assessments is below the threshold. In other words, the majority voting rule on the decision is equivalent to considering that the group's assessment is the median of the individual assessments. Hence, majority voting is in the race. What are the other challengers? A standard way to aggregate probabilities is to make averages. The linear average is defined as $\sum_k a_k$, and it can be generalized with weights $\omega_k \geq 0$ and $\sum_k \omega_k = 1$, as $\sum_k \omega_k a_k$, to take into account unequal degrees of expertise on the question.[3] Other averages are the geometric average or the harmonic average. Our problem is to determine which probability aggregation rule, followed by the threshold, is the best one in our problem. It is easy to see that these various probability aggregation rules can give different binary decisions for the group.[4]

Pooling probability functions has been studied for several years in the theory of probability aggregation (for surveys, cf. Dietrich and List forth., Martini and Sprenger forth., section 3). Can its results be used to select the best aggregation rule in our problem? I shall argue that unfortunately no. The framework of probability aggregation adopts an axiomatic method: it starts by stating several axioms which appear as desirable properties for the pooling function and then studies which function or aggregation rule, if any, satisfies them. The axioms considered in Dietrich and List's survey can be expressed in our case as:

- **Independence**: the group's probability a only depends on the individual probabilities a_k.
- **Unanimity preservation**: if all agents' probabilities a_k are the same, then the group's probability a is this one too.
- Three **Bayesian axioms**: if some information is learned by all individuals, then the group's decision changes by conditionalization on that event.

[3] It is akin to the iterated Lehrer-Wagner model which, starting from respect weights agent have to one another, provides a single probability for the group. However, the iterated Lehrer-Wagner model, and even more its normative interpretation, have been subjected to many criticisms (for a survey, cf. e.g. Martini and Sprenger forth. section 4). As a descriptive model, it is not useful for the present discussion.

[4] Consider for instance the median and the linear average, with three experts with $a_1 = a_2 = 0.04$, $a_3 = 0.10$, and $t = 0.05$. A majority voting on the decision gives a "true" as two experts on three assess the risk to be below the threshold. The linear average (with equal weights) is 0.06, which is higher than t, so this gives a "false".

The Independence axiom is automatically satisfied here, because our problem contains only one true/false answer, and there is no other probability on which a could depend. The three Bayesian axioms make sense in cases where the expert panel learns new information. In our problem, however, an extensive discussion has already taken place so no agent learns new information anymore, and the expert panel is not making any new inquiry. So the Bayesian axioms are not relevant in our case, and only the Unanimity preservation axiom expresses a desirable property for the aggregation rule.

An essential point to note is that a very large number of aggregation rules satisfy this axiom: the median, linear averaging, geometric averaging, and so on — actually, any convex function of the a_k. This illustrates the fact that a classical uniqueness result from the probability aggregation literature does not hold anymore: the well-known theorem by McConway 1981 and Wagner 1982, which states that linear averaging functions are the *only* independent and unanimity-preserving functions. The reason is that the theorem requires a set of at least three events, whereas our problem only considers two — e.g. the product is risky, with probability a_k, and the product is not risk, with probability $1 - a_k$. Considering a simpler agenda has widened the set of suitable aggregation rules, and no theoretical result from the literature can be used to pick the best one. More generally, the uniqueness and impossibility results from the theory of probability aggregation are useless for our problem. So, how scientific expert panels should aggregate risk assessments is not a simple problem that can be solved straightforwardly with the existing literature, which has focused on general problems with complex agendas, and has thus neglected more specific yet important questions. In the next section, I discuss other desiderata or axioms that we would like to impose on the aggregation rule.

3 Robustness Matters

Scientific risk assessment is supposed to meet some standards of reliability and objectivity, and the aggregation of these assessments should follow alike standards. In this spirit, I now introduce several new requirements for our aggregation rule. The aggregation rule should be sensitive to the right features of our problem, and not to the parasitic ones. It should favor objective features at the detriment of idiosyncrasies or unwanted values (for an analysis of the concept of objectivity, cf. Douglas 2004 — I refer to some of her distinctions below). In other words, the aggregation rule should be robust to some changes that we regard as irrelevant. In this section, I defend three dimensions of robustness that should be taken into account: the risk metrics, the level of detail, and the presence of strategical agents.

Several probability aggregation rules can be considered: linear averaging, geometric averaging, harmonic averaging, among others. As the forthcoming robustness discussion is similar for all the various averagings, I shall simplify it and consider only linear averaging, which shall be contrasted with the median. $\mathcal{R}_a$ denotes the aggregation rule that compares the threshold with the linear average (which thus stands for other averages), and $\mathcal{R}_m$ the aggregation rule that compares the threshold with the median of the individual assessments (which is equivalent to a majority vote on the decision itself).

3.1 Metrics

The formal model I have introduced relies on a quantitative scale — a and t are given numerical values in $[0, 1]$. How is this scale defined in real cases? My talking about probabilities has been only a matter of simplicity given the reduction of the problem to the $[0, 1]$ interval, and typical cases do not bear on well-defined probabilities or explicit scales. For instance, a standard question posed at an FDA advisory committee is "Does the overall risk versus benefit profile for X support marketing in the US ?"[5]. This question supposes that experts identify the risk versus benefit profile, and determine the value of the threshold under which a marketing is warranted. This can be done in a number of ways, and these are essentially value-laden questions[6] — what is acceptable or not has to do with extra-scientific values, and may also reflect the fact that an expert is risk-averse or risk-seeking. Overall, it makes sense to suppose that both the metrics scale and the threshold depend on the experts. Conversely, as the aggregation procedure is supposed to take place when the experts have extensively discussed, one can make the simplifying assumption that the same facts are known to all, and thus that the risk assessment is the same for all. In that way, our model actually applies in the setting in which a is common to all experts, but each has her own threshold t_k. The fact that the quantitative risk scale is not uniquely defined can be approached from a mathematical viewpoint: any scale can be reparametrized by applying any continuous bijection from $[0, 1]$ to $[0, 1]$, such as $x \mapsto x^2$.

These points make a hard time for the rule $\mathcal{R}_a$ (and other non linear averagings). First, from a practical viewpoint, the dependence of the risk scale metrics on the expert prevents the use of rules which take as inputs the numerical values of the risk assessments or of the threshold. For instance, is it even possible for a chairman to ask her colleagues "Please tell me your overall risk versus benefit acceptability threshold"

[5]Cf. Urfalino and Costa (2015, p.183).

[6]On the role of values in science more generally, and a critic of the value-free ideal, cf. Douglas 2009.

(or assessment), given that each expert may have her own scale? The rule $\mathcal{R}_m$, as it is equivalent to majority voting, needs not rely on input individual numerical values, and is thus safe from this criticism. Second, even if these practical difficulties could be overcome, some theoretical difficulties remain. Suppose a common scale has been adopted so that all experts can express their t_k. An aggregation rule that depends on the metrics of that common scale can give different outcomes according to the scale employed, as shown in Table 1. This dependence is a problem: which common scale should be chosen? (This is another aggregation problem!) Note that a variant of this problem exists even with a well-defined probability scale. For instance, let A be the event that a certain risk (e.g. carcinogenic substances in food) is responsible for more than 10 cases of cancer in 100,000 people during 1 year. The experts estimate the probability of A, $p(A)$. Consider now A' the event that the risk is responsible for more than 10 cases of cancer in 100,000 people during 10 years. Call $p(A')$ its probability. If the cancer cases are independent along the years, then $p(A') = 1 - (1 - p(A))^{10}$. Because the relation between $p(A)$ and $p(A')$ is not linear, taking the linear average of the experts assessments on A, and transforming it into an assessment on A', or taking the linear average of the experts assessments on A', does not give the same result. Which event A or A' is the more "natural" is not clear, and so much more for the right risk group assessment.

This gives good reasons to consider the following requirement: the aggregation rule should be insensitive to the metrics used to describe the problem, i.e. the assessment and the threshold. What should matter is just the relative position of the a and t_k, not their distance which can be due to some idiosyncratic value-laden judgments. This is requiring that the aggregation rule is more objective, under the sense of value-neutral objectivity as characterized by Douglas (2004, p. 460), which does not mean "free from all value influence" (as judging whether a risk benefit ratio is lower enough is bound to involve a value judgment), but takes a position "that is balanced or neutral with respect to a spectrum of values" (here, the balance is reached by taking into account only relative positions). The metrics robustness excludes the rule $\mathcal{R}_a$ which employs a linear average — Table 1 has shown a counter-

	t_1	t_2	t_3	a	Average t	$\mathcal{R}_a$	$\mathcal{R}_m$
x scale	.01	.01	.1	.05	.04	**False**	False
x^2 scale	0.0001	0.0001	.01	0.0025	0.0034	**True**	False

Table 1: Example in which the rule $\mathcal{R}_a$ gives different answers depending on the scale. The three experts have different thresholds t_k and a common risk assessment a.

example — but not $\mathcal{R}_m$ which relies on the median.[7]

3.2 Level of detail

Another argument for an aggregation rule that does not rely on a specific metrics comes from considerations of the level of detail in which the problem is described. So far, a continuous scale has been assumed, with numerical assessments in $[0, 1]$. Numerical discrete scales could also be used or even qualitative assessments only — it corresponds to decisions under uncertainty and not under risk. Consider for instance the case of the well-known IPCC Assessment Reports, that formulate a synthesis of existing scientific knowledge on climate change issues. The reports use a standardized vocabulary to express uncertainties, with several scales: some are qualitative (e.g. low/medium/high), others are quantitative (and use probabilities).[8] The historical trend has been to use more quantitative scales and less qualitative scales, but the latter have the advantage of being easily understandable by non-technical audiences, and thus should continue to be used in the future. Some qualitative and quantitative scales are in an explicit correspondence, as illustrated on Table 2. Writing an IPCC report involves synthesizing large amounts of scientific literature, so co-authors of a chapter may have different beliefs on the uncertainties associated with a finding. Whether they express their beliefs on a qualitative or on a quantitative scale, the way their beliefs are aggregated should be smooth and not vary abruptly (some very precise yet qualitative scales are conceivable), all the more than some explicit correspondence exist (Table 2). This is also a question of historically

[7] The comparability of scales is also discussed in Risse's (2004) political philosophy work, who also takes it as an argument for majority voting.

[8] Cf. e.g. the last report of the Working Group I, Stocker et al (2013, p. 138-142).

Term	Likelihood of the Outcome
Virtually certain	99–100 % probability
Very likely	90–100% probability
Likely	66–100% probability
About as likely as not	33–66% probability
Unlikely	0–33% probability
Very unlikely	0–10% probability
Exceptionally unlikely	0–1% probability

Table 2: Likelihood terms associated with outcomes used in the Fifth Assessment Report of the IPCC (Stocker et al 2013, p. 142).

consistency when switching from qualitative to quantitative scales.[9] Thus, a sound requirement is that the aggregation rule extends to formulations with discrete and qualitative scales. As the average of non-numerical and qualitative values is not defined, $\mathcal{R}_a$ does not satisfy this requirement. The median is defined on any kind of scale, and $\mathcal{R}_m$ satisfies the requirement. So only $\mathcal{R}_m$ is robust for the level of detail.

3.3 Bias and strategical votes

Not all experts are moved by epistemic goals only, and conflicts of interests can arise. For instance, numerous controversies have surrounded the FDA advisory committees along the years (Urfalino and Costa 2015, p. 168-169.) If a better selection of experts may be the solution, the decision rule used in the expert panel can also reduce the impact of bias agents.[10] With $\mathcal{R}_a$, an expert can strategically express a much lower risk of a medicine to influence the group's average — with a threshold at 10 %, she might express 0.1% instead of just 9%. The aggregation rule should be insensitive to such a strategical vote manipulation, and this is all the more important as the biased agent may have already influenced other agents during the preceding discussion. $\mathcal{R}_m$ is clearly robust in this sense, as an agent has the same influence whether her probability is just below the threshold or close to 0. This is not so for $\mathcal{R}_a$. This robustness requirement also makes the aggregation rule more objective, in the sense of detached objectivity (Douglas 2004, p. 459): one's personal values (allegiance to a firm) should not prevail on evidence (e.g. that the probability is 9%, as above).

Overall, the three robustness requirements considered here clearly favor $\mathcal{R}_m$ over $\mathcal{R}_a$. This provides a substantial justification for the traditional democratic rule in expert panels confronted with a binary decision. This result is a real departure from probability aggregation theory, in which linear averaging is justified on solid grounds. Narrowing the agenda and introducing a threshold has changed the solution to the aggregation problem.

[9] One may object that in the IPCC case the co-authors aggregate beliefs without a threshold comparison for a binary decision. Actually, thresholds are implicit: a finding which confidence is too low may not be mentioned. Anyway, the IPCC example can be seen as a mere illustration of the level of detail problem.

[10] Biased and extremist agents have been much studied in the literature of opinion dynamics (cf. for instance in Lorenz's 2007 survey), but not so in the literature of opinion aggregation.

4 Reasons

So far, a simplified model of scientific expert panels has been considered, one in which the group is asked to give a binary decision. As argued, the first step in justifying that decision consists for the panel to have a belief on the risk assessment, which is given by the median of the individual assessments in the case of $\mathcal{R}_m$. However, expert panels are often asked to provide a deeper justification. The question then arises of how the panel should aggregate its members views on this justification. In this section, I propose a novel but simple model for individual numerical assessment justification, in line with my previous threshold model.

Perhaps the most typical interpretation of the risk assessment a is that of a (subjective) probability. Suppose this probability is determined by m independent factors ($m \geq 2$). For instance, the risk associated with a medicine comes from m unrelated secondary effects. Then a is the probability that at least one risk factor triggers:

$$a = 1 - \prod_{j=1}^{m}(1 - a_j). \tag{1}$$

Each expert k is supposed to have her own assessment of each factor $a_{k,j}$ ($j = 1, \ldots, m$). Our problem is then to aggregate the $n \times m$ matrix of probabilities $a_{k,j}$, and to compare that result with the threshold.

As the m factors are independent, a sound requirement is to aggregate the individual assessments on them separately. How should that be done? Adapting the arguments from the previous section, one is lead to the conclusion that the panel should take the *median* of the individual assessments for each factor. However, there is a fundamental limitation to this, due to the previously mentioned theorem by McConway and Wagner's (cf. Section 2). Here is why. Requiring as above that the aggregation proceeds on each factor independently is just requiring the classical independence axiom. Another legitimate requirement is the classical axiom of unanimity preservation: if all experts agree on the risk assessment for one factor, then the panel should take this value as its own. As $m \geq 2$, all the conditions of the theorem by McConway and Wagner are fulfilled[11], so its conclusion apply: the only probability aggregation rule on the set of factors and on the overall decision is linear averaging. This reveals that, if groups use the median to determine both the independence factors' values and the overall risk (according to the above results), then it does not give a probability function and inconsistencies can arise. Table 3

[11]Each of the $m \geq 2$ factors can be triggered or not, so there are at least 4 events, which is higher than the 3 required in the theorem.

gives such an example. In other words, asking the expert panel to take stands on the reasons for its majority decision can lead it to change its decision.

Does it mean that our robustness defense of the median should be discarded? Not necessarily. The theorem by McConway and Wagner assumes that the experts aggregate their views *both* on the independent factors and on the overall risk assessment. But one can have the experts aggregate their views on the independent factors only. The overall risk assessment is then computed according to Equation 1, and the final decision is logically obtained from a comparison between this value and the threshold. In that way, experts do not vote on the final decision directly. This decision rule is a so-called premise-based rule.[12] Then, the linearity result of McConway and Wagner does not apply any more. The robustness considerations from the previous section do apply at the level of independent factors, and they recommend that the group takes the median of the individual assessments.

The present model of factors has assumed that there exists some common numerical scale, so that taking the median of individual assessments makes sense. However, the previous section has in part argued that such a scale may not always exist. In these cases, the present model of independent factors cannot apply. The theory of judgment aggregation offers a general framework for the aggregation of non-numerical reasons or justifications, with true/false beliefs (for reviews, cf. List 2012, Martini and Sprenger forth.). Applying in detail this framework to our problem of scientific justification would require another paper. A general result from this literature, however, is the discursive dilemma: majority voting on a set of true/false beliefs related in a logical way (here: reasons for the decision) may generate inconsistent collective judgments. This echoes our own finding about the median, which corresponds to majority voting in case of a threshold comparison. So whatever

[12] On this strategy more generally, see Cooke (1991), Bovens and Rabinowicz (2006), Hartmann and Sprenger (2012). Another solution to our problem could be the conclusion-based rule, i.e. aggregate only the views on the conclusion, but this is just like the previous section that we are trying to surpass.

Risk aspect	a_1	a_2	$a = 1 - (1 - a_1) \cdot (1 - a_2)$
Agent #1	0.01	0.01	0.0199
Agent #2	0.02	0.01	0.0298
Agent #3	0.01	0.02	0.0298
Median	0.01	0.01	0.0199 or 0.0298 ?

Table 3: A case in which the rule of the median can lead to inconsistencies. With a threshold at e.g. 0.025, the group's decision could be either true or false.

the scale, majority voting on all parts of the question is in great difficulty, and a premise-based solution should be adopted.

5 Conclusion

This paper has investigated the rationale for the majority rule that is often used in scientific expert panels, when dissent persists after discussion, and has looked for the best decision rule in this context. To this end, I have introduced a threshold probability model for individual decisions. Three main points have been shown in the paper: (1) the standard framework of probability aggregation is unable to solve our problem of risk aggregation. (2) robustness considerations clearly favor majority voting on the decision, i.e. comparing the threshold to the median of the individual risk assessments. (The robustness axioms I have advocated, which have been designed from considerations on scientific expert panel, could in return inspire social choice theory). (3) when a justification of the panel's decision is looked for, the median rule (corresponding to majority voting) can lead to inconsistencies. The promising route is to have the group aggregate on the reasons level, not on the final decision one. This should encourage scientific expert panels to divide questions from a logical viewpoints, and to take decisions on sub-problems instead of voting on the final decision directly. Current practices in advisory committees of the FDA and of the EMA could evolve in this respect. However, these claims have only been shown in quite simple and idealized models of decision-making. Future work is needed to investigate other models. These preliminary results have nonetheless cast some serious doubts on the majority voting rule only applied on the final decision.

Note finally the generality of the proposed model, which goes well beyond scientific expertise: the a and t variables can be interpreted as degrees of beliefs or as utility measures, within an epistemology or an economy framework.

References

Bovens, Luc and Wlodek Rabinowicz. 2006. "Democratic answers to complex questions. An epistemic perspective". *Synthese* 150: 131-153.

Cooke, Roger M. 1991. *Experts in Uncertainty. Opinion and Subjective Probability in Science*. Oxford University Press.

Dietrich, Franz and Christian List. Forthcoming. "Probabilistic Opinion Pooling". In *Oxford Handbook of Probability and Philosophy*, Oxford University Press.

Douglas, Heather E. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138:453-473.

Douglas, Heather E. 2009. *Science Policy and the Value-Free Ideal.* University of Pittsburgh Press.

Hartmann, Stephan and Jan Sprenger. 2012. "Judgment aggregation and the problem of tracking the truth." *Synthese* 187:209-221.

Hauray, Boris and Philippe Urfalino. 2007. "Expertise scientifique et intérêts nationaux. L'évaluation européenne des médicaments 1965-2000". *Annales HSS* 2: 273-298.

List, Christian. 2012. "The theory of judgment aggregation: an introductory review." *Synthese* 187:179-207.

Lorenz, Jan. 2007. "Continuous Opinion Dynamics under Bounded Confidence: A Survey." *International Journal of Modern Physics C* 18, 1819.

Martini, Carlo and Jan Sprenger. Forthcoming. "Opinion aggregation and individual expertise." In *Scientific collaboration and collective knowledge*, ed. by T. Boyer-Kassem, C. Mayo-Wilson and M. Weisberg, Oxford University Press.

McConway, Kevin J. 1981. "Marginalization and Linear Opinion Pools." *Journal of the American Statistical Association* 76(374): 410-414.

Risse, Mathias. 2004. "Arguing for Majority Rule". *The Journal of Political Philosophy* 12(1): 41-64.

Stocker Thomas .F. et al. 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Urfalino, Philippe and Pascaline Costa. 2015. "Secret-Public Voting in FDA Advisory Committees." In *Secrecy and Publicity in Votes and Debates*, ed. Jon Elster, 165-194. Cambridge University Press.

Wagner, Carl. 1982. "Allocation, Lehrer models, and the consensus of probabilities." *Theory and Decision* 14: 207-220.