Opinion Aggregation and Individual Expertise

Carlo Martini* Jan Sprenger†

March 20, 2015

Abstract

Group judgments are often influenced by their members' individual expertise. It is less clear, though, how individual expertise should affect the group judgments. This paper surveys a wide range of formal models of opinion aggregation and group judgment: models where all group members have the same impact on the group judgment, models that take into account differences in individual accuracy, and models where group members revise their beliefs as a function of their mutual respect. The scope of these models covers the aggregation of propositional attitudes, probability functions, and numerical estimates. By comparing these different kinds of models and contrasting them with findings in psychology, management science and the expert judgment literature, we gain a better understanding of the role of expertise in group agency, both from a theoretical and an empirical perspective.

Contents

*Contact information: Academy of Finland Centre of Excellence in the Philosophy of the Social Sciences, Department of Political and Economic Studies, P.O. Box 24, University of Helsinki, 00014, Helsinki (Finland). Webpage: `http://www.martinicarlo.net`. Email: carlo.martini@helsinki.fi.

†Contact information: Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Webpage: `http://www.laeuferpaar.de`. Email: j.sprenger@uvt.nl.

1 Introduction

Members of a group are often in **disagreement** with each other. Analysts at Apple or Samsung come up with different estimates of how often a newly developed cell phone will be sold. Conservation biologists disagree on the population size of an endangered species. Researchers at the European Central Bank cannot find a consensus on the merits and drawbacks of a particular monetary policy. In all these cases, disagreement need not be a bad thing. What is more, it can be seen as epistemically and socially desirable. Armstrong (2001) and Page (2007) argue, among others, that the diversity of opinion characteristic of disagreement can act as an antidote to groupthink and foster the development of alternative approaches whose pursuit may be more fruitful than following well-trodden paths.

On the other hand, disagreement can block the formation of group judgment and delay important decisions. To stay with one of the above examples, the European Central Bank needs to come up with *some* decision on whether to ease or to tighten monetary policy. Also in science, it seems that a certain degree of consensus is necessary for scientific progress and for conducting "normal science" (Kuhn 1962, 1977). So how should groups **aggregate (or pool) the opinions of their members**?

This question has no clear-cut, unique answer. One reason is the diversity of contexts where groups aggregate individual opinions, and the different goals they aim at. Another reason is the diversity of criteria for evaluating aggregation procedures, e.g., epistemic and social ones. In the epistemic perspective, opinion aggregation procedures are primarily assessed according to their veracity, that is, their success at tracking the truth. In the social perspective, they are judged on different criteria: Was the opinion of every individual duly taken into account? Is the final result acceptable to all group members? Did it emerge from a procedure that everybody consented to? And so on. To give a crude example, a dictatorship is usually ruled out in the social perspective, but it may be acceptable from

an epistemic perspective, if that individual is by far the most competent group member.

There are two principled ways of aggregating opinions and resolving disagreements: with and without **belief revision on an individual level**. The main part of the paper is devoted to models of opinion aggregation that do not involve belief revision. Such models describe in a formal framework how opinion aggregation procedures should (or factually do) work.[1] However, we will also contrast them with consensual decision-making where individuals revise their beliefs and eventually agree on the subject matter of disagreement. Typical driving factors of such procedures are deliberation and considerations of power and mutual respect.

It is useful to divide opinion aggregation models that do not involve deliberation or belief revision into two categories. In the first category are **egalitarian models**—a term supposed to capture the fact that no individual has a special or privileged position. They are particularly important in contexts where it is hard to argue for giving special weight to a specific individual, e.g., because relevant expertise is hard to elicit. Sometimes, it is even one of the constraints on an aggregation procedure that the opinions be "anonymous", that is, that the final outcome (the group opinion) do not depend on the individual agents that submitted a view. This aggregation procedure is acting like the famous allegory of justice: It weighs opinions impartially and regardless of the agent's identities.

The use of egalitarian models may have epistemic, procedural or pragmatic reasons. For example, the equality of individual votes in elections or referenda—a particular case of social opinion aggregation—has constitutional status in most democracies. Therefore, choosing an egalitarian procedure may be pragmatically required. Hence, it comes at no surprise that egalitarian aggregation procedures, such as various forms of majority or plurality voting, play a major role in political processes.

The second category of aggregation models—which we will call **differential models**—give differential weights to the individual group members. This procedure can have psychological as well as epistemic motivations: Firstly, "fringe opinions" — i.e. those that are far away from the group average — may receive low weight precisely because they appear to be wide off the mark, and nobody else is willing to take them seriously. Secondly, when the task is *intellective*, that is, when it involves a high level of demonstrability, some group members may be more competent than others. In

[1] We use "procedure" as a generic word for ways to aggregate opinions and to resolve disagreements, whereas "model" denotes a particular formal framework for describing these procedures.

such tasks, the opinions of the **experts** might receive higher consideration in determining the group view. Here we think of experts as the more competent group members, that is, those whose opinion coincides more often with the truth, or contains the smallest error. Thirdly, there are group members that may, because of their position or their appearance, receive more respect or esteem than others in the peer-group, and it may be procedurally or pragmatically necessary to endow them with a higher weight than their peers.

Which of all these procedures is applied in practice is a context-sensitive issue: For instance, in a meeting of the heads of state of the European Union on foreign policy or the EU budget, the primary goal may consist in finding a consensual position, and that need not necessarily involve egalitarian decision-making. Also when the epistemic accuracy of the group judgment has the highest priority, e.g., in scientific reasoning, it may be rational to weigh differentially and to defer to experts. More generally, List (2005) distinguishes between two challenges for group judgments: the "rationality challenge" of endorsing a consistent collective judgment on an interconnected agenda of propositions, and the "knowledge challenge", that is, to track truth in these collective judgments. Both challenges will be discussed in this paper.

Finally, we should distinguish between weighting *strategies* and weighting *schemes*. A strategy is general: In the classification presented in this paper, it is either egalitarian or differential. A scheme is specific: It is clear that once a differential weighting strategy has been selected, a specific set of weights (a scheme) is needed. Perhaps contrary to our intuitions, egalitarian aggregation procedures can allow for different weighting schemes. Equal weights views in the epistemology of disagreement literature are often vague on the issue of which specific function aggregates the individual opinions; for example, the "equal-weight view" is sometimes interpreted in a Bayesian scheme (Elga 2007) and other times as linear averaging (Christensen 2009).

In sum, this paper sketches a rough map of diverse approaches to aggregating group opinions that matter for science, politics and social decision-making more generally. Particular attention will be devoted to differential models, and among them, to models where an agent's weight is determined by his or her expertise in the subject area. These models are contrasted with egalitarian or power-based opinion aggregation. Notably, we omit the problem of information aggregation and assume that all agents are on a par with respect to the available information. Possi-

ble differences between them are thus a matter of different judgmental prowess.

The two following sections are devoted to the problem of group rationality in opinion aggregation: how should a group aggregate individual opinions if constrained by a set of epistemic or social requirements? While Section 2 deals with aggregating **binary propositional attitudes**, that is, yes-no judgments, Section 3 investigates the aggregation of **graded attitudes**, and in particular, probability distributions. Sections 4 and 5 both describe the problem of aggregating **numerical estimates**, but from different perspectives. First, we examine philosophical and empirical motivations for **consensual belief revision** in the light of disagreement and review a major formal model of such a procedure, the Lehrer-Wagner model. Then, we conduct an analysis of the epistemic benefits of differential weighting as opposed to straight averaging. Section 6 gives an overview of empirical findings on expertise, e.g., the identification and ranking of experts, and relates them to the formal results. Section 7 concludes.

2 Judgment Aggregation

Judgment aggregation is a recent field at the intersection of philosophy and economics. It deals with the aggregation of individual judgments on an interconnected set of propositions into a coherent group judgment and asks what kind of aggregation procedure rationality requires us to choose. For our purposes, judgment aggregation is particularly interesting because procedural constraints may require democratic decision-making, where social and epistemic features of individual agents are not taken into account. Judgment aggregation procedures are typically egalitarian: no single person obtains more weight than others because of his/her standing in the group or his/her competence. Studying judgment aggregation, and its problems and methods, also yields a better understanding of the advantages and drawbacks of differential weighting in opinion aggregation.

Notably, it has been shown that the classical economic problem of preference aggregation can be embedded into the problem of judgment aggregation (Dietrich and List 2008; Grossi 2009). Classical results such as Arrow's impossibility theorem can be represented as impossibility theorems for aggregating judgments on logically interconnected propositions. This increases the generality and relevance of theoretical results on judgment

aggregation.

Let us now formulate the classical problem of judgment aggregation. Assume that N agents are supposed to aggregate their binary judgments on an agenda of propositions $X = \{A_1, \ldots, A_m\}$ where the A_k can be logically interconnected. Call J_i the judgment set of agent i. Then we can ask ourselves which kind of condition should be satisfied by an aggregation function $F : \mathcal{J}^n \to \mathcal{J}$ that maps the individual agents' judgment sets to a group judgment set $F(J_1, \ldots, J_n)$.

Among the most popular conditions are:

Universal Domain Any combination of consistent individual judgments is in the domain of F.

Collective Rationality $F(J_1, \ldots, J_n)$ is a consistent and complete collective judgment set on F.

Anonymity For any two profiles $(J_1, \ldots, J_n)$ and $(J'_1, \ldots, J'_n)$ which are permutations of each other, $F(J_1, \ldots, J_n) = F(J'_1, \ldots, J'_n)$.

The first two conditions, Universal Domain and Collective Rationality, express that we are looking for a general aggregation procedure that outputs consistent group judgments. Anonymity articulates the egalitarian intuition: the group judgment set is invariant with respect to the "position" of an agent in the group. This is evidently inspired by democratic decision-making where votes are cast anonymously and cannot be traced back to the voter. As a consequence, the idea of experts who have special weight in the group is eliminated in favor of having a genuinely egalitarian decision model.

For judgment aggregation, this natural idea has an awkward consequence as soon as another condition, **Systematicity**, is added. That condition requires, informally, (i) that the group judgment on each proposition in the agenda depend only on individual judgments on that particular proposition, (ii) that the same aggregation criterion be used for each single proposition. In a much-cited paper, List and Pettit show that the above four criteria are jointly incompatible with each other.

Theorem 1 (Classic Impossibility Result for Judgment Aggregation) *There is no judgment aggregation function that satisfies Universal Domain, Collective Rationality, Anonymity, and Systematicity.*

(List and Pettit 2002)

Group Member	p	$p \to q$	q
Alice	Yes	Yes	Yes
Bob	Yes	No	No
Carol	No	Yes	No
Majority Judgment	Yes	Yes	No

Table 1: The discursive dilemma for majority voting illustrates the impossibility result by List and Pettit.

This impossibility result has subsequently been extended and generalized—see List and Puppe (2009) and List (2012) for introductory reviews. A classic instance of List and Pettit's impossibility result is the **discursive dilemma** for majority voting. Assume that a group of three members, Alice, Bob and Carol, have to vote on propositions p, $p \to q$ and q. If they agree on majority voting as a judgment aggregation procedure for all three propositions, their consistent individual judgments may yield an inconsistent collective judgment. See Table 1 for an example.

This means that we have to part with at least one of the four above conditions—see List (2005) for an extended discussion. Giving up Universal Domain seems to make things too easy by restricting the types of disagreements that may occur in practice. Why shouldn't Alice, Bob and Carol be allowed to have all kinds of consistent individual judgments on the agenda $\{p, p \to q, q\}$?

Abandoning Collective Rationality may look counterintuitive, but need not be a bad idea. Suspending judgments on *some* of the propositions, or choosing context-sensitive aggregation procedures with a supermajority threshold may avoid inconsistencies. This move may also reflect that diverse aggregation problems demand diverse solutions. Still, it would be a more principled option to dispute Systematicity or Anonymity instead.

Indeed, Systematicity may appear unreasonably strong as soon as one realizes that propositions in an agenda are often interconnected: no judgment is entirely independent of judgments on the other propositions. In addition, for some propositions, e.g., those where an error has severe practical consequences, it may be reasonable to use a different quorum than for other propositions in the agenda. Such a move would contradict Systematicity. Still, it is a delicate issue to prioritize some of the propositions over others, and good context-sensitive reasons need to be given. An epistemic justification for premise-based judgment aggregation, that is, judgment aggregation on the *reasons* for a group decision, has been provided

by Bovens and Rabinowicz (2006), Hartmann, Pigozzi and Sprenger (2010) and Hartmann and Sprenger (2012).

In favor of abandoning Anonymity, it could be said that procedural uniformity (as encoded in Systematicity) is a big practical asset. Also, some agents may possess a higher level of expertise than others. All this motivates the abandonment of Anonymity and the investigation of differential models where systematicity-like conditions can be maintained. We will examine them in the next section. However, practical constraints (e.g., democratic decision-making) often require that Anonymity be endorsed. Therefore, a solution among these lines can only be partial.

Let us now turn to the aggregation of fine-graded epistemic attitudes, such as probabilistic degrees of belief. Perhaps the impossibility results for judgment aggregation are just an artifact of the binary aggregation setting and will disappear in the more expressive probabilistic framework?

3 Probability Aggregation

The problem of **probability aggregation** is to reconcile probability measures over a σ-algebra $\mathcal{A}$ into a single (group) probability measure. These probabilities can be naturally interpreted as representing epistemic attitudes, that is, as individual and collective degrees of belief over the propositions in the algebra. Formally, we are looking for an aggregation function $F : \mathcal{P}^n \to \mathcal{P}$ that maps individual probability measures over $\mathcal{A}$, $(p_1, \ldots, p_n)$, to a group probability measure p^*.

A natural constraint on probability aggregation is **Convexity**, stating that the group probability of a proposition A, $p^*(A)$, should lie in between the minimum and the maximum of the individual probabilities $(p_1(A), \ldots, p_n(A))$. Therefore, some form of averaging (e.g. arithmetic or geometric, with or without weights) appears to be the natural solution to the problem of reconciling probability distributions.

Another natural constraint is the

Strong Setwise Function Property (SSFP) There is a function $g : [0,1]^n \to [0,1]$ such that for any event $A \in \mathcal{A}$:

$$F(p_1, \ldots, p_N)(A) = g(p_1(A), \ldots, p_N(A))$$

This condition requires, similar to Systematicity in judgment aggregation, that the group probability of any proposition A only depend on the individual probabilities of A, and be screened off from other propositions.

SSFP directly yields the

Zero Preservation Property (ZPP) If for a proposition $A \in \mathcal{A}$ and all group members $\{1, \ldots, N\}$, $p_1(A) = \ldots = p_N(A) = 0$, then also $F(p_1, \ldots, p_N)(A) = 0$.

ZPP expresses the very natural idea that if every agent considers an event impossible (probability zero), then also the group should find it impossible.

As shown independently by McConway (1981) and Wagner (1982), any probability aggregation function that satisfies SSFP is a linear aggregation rule:

Theorem 2 (McConway 1981, Wagner 1982) *- Any probability aggregation function $F : \mathcal{P}^n \to \mathcal{P}$ over an algebra $\mathcal{A}$ that satisfies SSFP is of the form*

$$F(p_1, \ldots, p_N)(A) = \sum_{j=1}^{n} \omega_j p_j(A) \tag{1}$$

for some weights $\omega_1, \ldots, \omega_n \in [0, 1]$ such that $\sum_{j=1}^{n} \omega_j = 1$.

In other words, probability aggregation in agreement with SSFP reduces to a linear average of the individual probabilities. Note that the weights need not be equal to each other, as a purely egalitarian model would require.

This consequence of SSFP is very elegant, but it leads into trouble—like the consequences of its judgment aggregation counterpart Systematicity. Consider the following property:

Independence Preservation Whenever for two propositions $A, B \in \mathcal{A}$, we have $p_j(A \wedge B) = p_j(A) \cdot p_j(B)$ for all $1 \leq j \leq n$, then also

$$F(p_1, \ldots, p_N)(A \wedge B) = F(p_1, \ldots, p_N)(A) \cdot F(p_1, \ldots, p_N)(B)$$

In other words, if all group members agree that two propositions are probabilistically independent, then this independence should be preserved in the group judgment. Unfortunately, this reasonable property is incompatible with SSFP:

Theorem 3 (Lehrer and Wagner 1983) *There is no non-dictatorial probability aggregation function $F : \mathcal{P}^n \to \mathcal{P}$ over an algebra $\mathcal{A}$ that satisfies SSFP as well as Indepedence Preservation.*

In addition, linear probability aggregation (that is, aggregation in agreement with SSFP) does not commute with Bayesian Conditionalization. And a natural rule that commutes with Conditionalization, like geometric averaging, fail to satisfy other desirable properties, such as Zero Preservation and Independence Preservation—see Brössel and Eder (2013) for an overview. Similar results hold, by the way, for the aggregation of *causal judgments*, that is, the aggregation of causal (in)dependency relations represented by directed acyclical graphs (Bradley, Dietrich and List 2014).

One of the responses to Lehrer and Wagner's impossibility results is to restrict SSFP to a subset of propositions (see Genest and Zidek (1986) for a review). This move is again parallel to the judgment aggregation literature where premise- and conclusion-based aggregation rules have been investigated. Another option is the adoption of a full Bayesian model for aggregating probability distributions (e.g., Lindley 1983).

The numerous impossibility results that exist in both domains—judgment aggregation and probability aggregation—suggest that the aggregation of epistemic attitudes on an interconnected set of propositions is just a very hard problem. It is also clear that the egalitarian presumption of judgment aggregation (Anonymity) is not the culprit for the impossibility results since similar results hold for linear aggregation procedures with differential weights. Rather, the results suggest that it is the logical connection between the propositions in the agenda that creates problems for "natural" aggregation procedures.

In the remainder, we therefore focus on aggregation procedures for a single proposition, or for estimators of a numerical quantity. This makes the impossibility results disappear, but it also leads to a shift in focus: we introduce models of differential opinion aggregation and compare their epistemic performance to egalitarian models.

4 Consensual Opinion Aggregation

When real groups make judgments or decisions, their members *interact* with each other: they exchange relevant information, put forward arguments and deliberate the reasons for a particular position. The previous sections did not take such interactions into account. And while it would go beyond the scope of this article to review the psychological and philosophical literature on group interactions, a particular phenomenon is relevant for our purposes: the tendency toward uniformity.

A long research tradition in social psychologists explores how groups

combine individual judgments, and which decision rule (e.g., simple majority, weighted majority, "truth wins", etc.) describes the group behavior best (e.g., Lorge and Solomon 1959). In a classic study on group judgments in intellective problems (that is, problems with a high level of demonstrability), Thomas and Fink (1961) compare three different models: an **independent model**, where the group reliability is just the probability that each group member has solved the problem, a **rational model**, where the group makes a correct judgment as soon as a single member is right, and a **consensus model**, which assumes the group's inclination toward uniformity. Using an arithmetically simple, but conceptually tricky mathematical problem, the authors find that the consensus model describes the outcomes better than the other two.

Presumably inspired by these findings, Davis (1973, 122) developed the influential Social Decision Scheme (SDS) model where probability distributions that describe individual preferences are transformed into a group probability distribution over the alternatives. That is, group members make an individual judgment in terms of a probability distribution which is subsequently transformed into a group judgment, by means of a matrix multiplication procedure. Extensions of this approach to the problem of combining numerical estimates have been provided by Davis (1996)'s Social Judgment Scheme, and Hinsz (1999)'s SDS-Q ("Q" standing for "quantitative") models. In these cases, weights may also be determined as a function of the *centrality* of an estimate and its distance to other estimates.

While these models are silent on the mental attitudes of the group members, they inspire the philosophical question of whether it is rational to revise one's beliefs or estimates in the light of disagreement with other group members. This question has received much attention in the recent epistemology literature on peer disagreement (e.g., Kelly 2005; Elga 2007), but it has also been applied to group judgments. In particular, philosophers have asked themselves whether it is possible to give rational foundations to consensual opinion aggregation.

The most popular idea is that **mutual respect** among the group members should prompt every group member to revise her initial opinion. This respect can be epistemically motivated (e.g., by realizing that the other group members are no less competent that oneself), but also reflect degrees of care or relations of social power, dependent on whether there is a matter of fact to the subject of disagreement. Conditional on such mutual respect, blending one's opinions with the opinions of the other group

members seems to be a requirement of rationality:

> One justification for aggregation is consistency, since refusing to aggregate is equivalent to assigning everyone else a weight of zero [...]. (Lehrer and Wagner 1981, 43)

In other words, refusing to blend one's opinions would amount to unjustified *dogmatism* (see also Lehrer 1976). This argument is, by the way, independent of the question of whether the aggregation procedure should be differential or egalitarian. It just motivates the view that group rationality need not be a question of choosing the right aggregation procedure at group level, but also a question of individual belief revision.

Among respect-based models of opinion aggregation, the Lehrer-Wagner model is most prominent. It was first developed as a descriptive mathematical model of group power relations in French (1956) and as a general model of consensus formation in DeGroot (1974). While DeGroot intended his model to be normative, his research focused on the fundamental mathematical properties of the model, leaving the interpretation, further elaboration, and philosophical justification to Lehrer and Wagner (1981).

The model tackles the problem of estimating a particular quantity x, from the individual estimates v_i of every group member i. This quantity x is normally thought of as objective and independent of the group members' cognitive states. The quantity x in dispute might, for instance, be the size of the population of an endangered species, or the number of rainy days in the Netherlands in 2014.

Lehrer and Wagner's central idea consists in ascribing the agents beliefs about each other's expertise, or in other words, mutual assignments of respect as epistemic agents on the issue at hand. Then, the w_{ij} describe the proportion to which j's opinion on the subject matter in question affects i's revised opinion. These mutual respect assignments are used to revise the original estimates of the quantity in question, and codified in an $N \times N$ matrix W (where N denotes the number of agents in the group):

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1N} \\ w_{21} & w_{22} & \dots & w_{2N} \\ \dots & \dots & \dots & \dots \\ w_{N1} & w_{N2} & \dots & w_{NN} \end{pmatrix}.$$

The values in each row are nonnegative and normalized so as to sum to 1: $\sum_{j=1}^{N} w_{ij} = 1$. Thus, the w_{ij} represent relative weights which the agents ascribe to themselves and to others when it comes to estimating the unknown value x. Then, W is multiplied with a vector v that contains

the agents' individual estimates of x, obtaining a novel updated value for v:

$$W \cdot v = \begin{pmatrix} w_{11}v_1 + w_{12}v_2 + \ldots + w_{1N}v_N \\ w_{21}v_1 + w_{22}v_2 + \ldots + w_{2N}v_N \\ \ldots \\ w_{N1}v_1 + w_{N2}v_2 + \ldots + w_{NN}v_N \end{pmatrix}.$$

In general, this procedure will not directly lead to consensus, since the entries of $W \cdot v$ differ: $(Wv)_i \neq (Wv)_j$. However, Lehrer and Wagner (1981) show that under very weak constraints, the sequence $(W^k)_{k\in\mathbb{N}}$ converges to a matrix W^∞ where all rows are identical, that is, where all agents agree on their relative weights. That is, when the procedure of averaging is repeated, the agents will finally achieve a consensus and not only agree on the factual subject matter, but also on the differential weight that each group member should obtain.

Is the Lehrer-Wagner model a *rational* model of disagreement resolution and group decision-making? On the pro side, it has been shown that the opinion changes of the agents can be represented as a form of Bayesian updating: agents exchange information and at every step, their degrees of belief are revised by conditioning on the pieces of information they receive from their peers (Geanakopoulos and Polemarchakis 1982; Romeijn and Roy 2015). So Lehrer-Wagner opinion aggregation can, at least in principle, be made compatible with Bayesian rationality standards.

On the con side, it has been argued that Lehrer and Wagner fail to provide an epistemic story of why we should change our opinions in this peculiar way. Suppose an agent determines and normalizes her respect weights for the other group members. Then, it remains opaque why we should choose a linear updating model rather than geometric weighting or another form of averaging (Martini, Sprenger and Colyvan 2013). Indeed, linear averaging is particularly sensitive to outliers in the individual estimates. If the group members determine the relative weights as a function of their mutual respect and independent of the submitted estimates, then extreme opinions will have an overly large impact on the group consensus.

Another worry is that respect-based differential weights may be caused by various forms of bias (Faust 1984; Trout 2009). When we assign weights based on mutual respect, biases easily distort the reasons for the weights assignment, so that there might be very little relation between the weights assigned and the objective, epistemic, weight that a certain opinion should receive. See the following section for a deeper analysis of this worry.

The most prominent objection, however, concerns the justification for *repeated* linear averaging. Lehrer and Wagner (1981) provide two stories why this might be rational. The first is a *temporal interpretation*: the disagreement after the first round of averaging is qualitatively similar to the initial disagreement. So the only way to avoid unjustified dogmatism is to repeat belief revision until consensus is reached. Wagner (1978) draws an analogy to sharing anonymous position papers after each round of aggregation: agents distribute their revised view and the reasons for that view among the other group members. This is actually similar to the famous Delphi method for structured forecasting developed by the RAND corporation (Helmer-Hirschberg 1967; Linstone and Turoff 1975): in that procedure, group members fill in questionnaires, comment on their reponses and subsequently receive a (filtered) summary of the opinions of the other group members. This procedure is repeated until consensus is eventually reached.

A concern with this interpretation is that second-order expertise does not play any role. One need not be an expert oneself in order to make an accurate judgment—as long as one is able to identify which group members are competent and which aren't. If the Lehrer-Wagner model strives to be a model of group rationality, such considerations should not be neglected. However, the second interpretation of the iterated matrix multiplication that Lehrer and Wagner provide—as an amalgamation of different orders of expertise—quickly becomes cognitiviely absurd since one cannot rationally believe that considerations of fourth-, fifth- or sixth-order expertise should be as important as first- and second-order expertise.

For more elaborate criticism of the Lehrer-Wqagner model, see Loewer and Laddaga (1985) and Martini, Sprenger and Colyvan (2013). The latter argue that the Lehrer-Wagner model should be understood as a model of social influence and not as a genuinely epistemic model. In the view of Martini, Sprenger and Colyvan, respect-based models of opinion aggregation are better suited for judgmental than for intellective tasks: that is, for making decisions on non-factual matters, or for aggregating preferences in a group of agents that respect each other, such as friends or colleagues. For modeling how individual *expertise* increases group accuracy, models where the weights are not determined by mutual respect, but by properly epistemic considerations may be more adequate.

5 An Epistemic Analysis of Differential Weighting

The previous section has investigated differential models of opinion aggregation where the individual weights are correlated with a group member's social influence or perceived expertise. In this section, we conduct an epistemic analysis of differential weighting procedures: which relations between objective performance indicators and the relative weights have to hold to improve group performance with respect to the natural benchmark, straight averaging? This also addresses a lacuna in the Lehrer-Wagner model where such a link between truth-tracking and specific weighting schemes is not provided.

A fundamental objection to differential weighting, that we address first, is the problem of identifying experts. This question gains urgency in the light of studies that show that in spite of modest correlations, there is a substantial gap between actual and perceived expertise (e.g., Trotman et al. 1983; Littlepage et al. 1995). However, expertise may also be recognized implicitly. In this case, experts tend to exert greater influence on the group judgment than non-experts (Bonner, Baumann and Dalal 2002). This is good news because it indicates that expert-dependent decision schemes may successfully operate without the ability to make explicit such a ranking. Indeed, research on intellective tasks such as *Mastermind* and letter-number-matching demonstrates that groups tend to use expert-weighted social decision schemes for such tasks, and achieve a performance that roughly corresponds to the best individual member (Bonner 2004; Baumann and Bonner 2004). This sounds modest, but actually, it is a substantial achievement if we do not know in advance who the experts are (see also Laughlin and Ellis 1986; Libby, Trotman and Zimmer 1987; Laughlin, Hatch et al. 2006; Bonner, Silito and Baumann 2007). In this context, it is notable that performance feedback does not substantially help the agents to recognize expertise and to improve performance. Anyway, these findings indicate that differential weighting can be epistemically beneficial in a variety of contexts, and that expert recognition is no practically impossible task. This brings us to the theoretical question of *how* we should weigh the experts.

A classical result in this area concerns the **aggregation of binary forecasts**. Assume that we have to predict whether it will rain on the next day. The agents are conceptualized as a group of independent forecasters with a certain probability p_i of getting the result right. How should these forecasts be combined? Nitzan and Parous (1982) and Shapley and Grofman (1984) show that if rain is a priori as likely as no rain, group accu-

racy is maximized by following a weighted majority voting rule where the weights of the agents are proportional to their logarithmic betting odds:

$$w_i \propto \log \frac{p_i}{1 - p_i} \tag{2}$$

This result is, in fact, closely related to Bayes' Theorem. However, it only solves the problem of **combining binary forecasts** and does not address the more general problem of **combining numerical estimates** of an unknown quantity μ. These estimates can correspond to the individual opinions of a group of agents, but also to the outputs of different mathematical models, e.g., different predictions for the extent of global warming. Our problem consists in finding a method of combining these estimates in an advantageous way, and stating general conditions for when taking into account individual expertise improves the group judgment.

We address this question through a simple statistical model developed by Klein and Sprenger (2015). Their work builds on analytical work in the forecasting and social psychology literature (Bates and Granger 1969; Hogarth 1978), following the approach of Einhorn, Hogarth and Klempner (1977). Because of its simplicity and generality, the Klein-Sprenger model is especially well suited for principled comparisons of egalitarian and differential opinion aggregation. It also stands in a venerable research tradition in social psychology: agents are modeled as independent signallers with a certain reliability (e.g., Zajonc and Smoke 1959). Such formal models are then used as a standard for gauging empirical findings, and they may indicate how information should be spread over the agents in order to optimize performance in a recall task or similar cognitive problems.

Klein and Sprenger model the group members' individual estimates X_i, $i \leq n$, as independent random variables that scatter around the true value $\mu = 0$ with zero bias and variance σ_i^2. No further distributional assumptions are made in order to preserve the generality of the analysis. The competence of an agent (or scientific model) is explicated as the degree of precision in estimating the true value. Then, the epistemic question about the epistemic benefits of differential weighting can be translated into a precise mathematical question:

> **Problem:** Which convex combination of estimates $\hat{\mu} = \sum_{i=1}^{n} c_i X_i$ should the agents choose in order to reduce expected square loss?

This mathematical question serves to identify a modeling target which is reasonably close to the actual problem and which we can use to study the

epistemic properties of differential weighting in opinion aggregation.

It is well-known that for any such estimate $\hat{\mu}$, the mean square error (MSE) can be calculated as

$$\begin{aligned} \mathrm{MSE}(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mu)^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} c_i X_i\right)^2\right] \\ &= \sum_{i=1}^{n} c_i^2 \mathbb{E}\left[X_i^2\right] + \sum_{i \neq j} c_i c_j \mathbb{E}[X_i]\mathbb{E}[X_j] \\ &= \sum_{i=1}^{n} c_i^2 \sigma_i^2 \end{aligned} \tag{3}$$

which is minimized by the following assignment of the c_i (cf. Lehrer and Wagner 1981, 139):

$$c_i^* = \left(\sum_{j=1}^{n} \frac{\sigma_i^2}{\sigma_j^2}\right)^{-1}. \tag{4}$$

The problem with these optimal weights is that each agent's individual expertise would have to be known in order to calculate them. They can also be quite extreme. Given all the biases that actual deliberation is loaded with, e.g., ascription of expertise due to professional reputation, age or gender, or bandwagon effects, it is unlikely that the agents succeed at unraveling the expertise of all other group members to such a precise degree (cf. Nadeau, Cloutier and Gray 1993; Armstrong 2001). In line with what has been said before, it is more realistic to expect that groups may be *qualitatively* competent at identifying experts, but not at determining the optimal weights.

Therefore the scope of the inquiry is widened:

> **Question:** Under which conditions will differentially weighted group judgments outperform the straight average?

A first answer is given by the following results where the differential weights preserve the expertise ranking (see Klein and Sprenger (2015) for all results and proofs):

Theorem 4 (First Baseline Result) *Let* $0 \leq c_1 \leq \ldots \leq c_n \leq 1$ *be the weights of the individual group members, that is,* $\sum_{i=1}^{n} c_i = 1$. *Further assume that for all* $i > j$*:*

$$1 \leq \frac{c_i}{c_j} \leq \frac{c_i^*}{c_j^*} \tag{5}$$

Then the differentially weighted estimator $\hat{\mu} := \sum_{i=1}^{n} c_i X_i$ *outperforms the straight average. That is,* $MSE(\hat{\mu}) \leq MSE(\bar{\mu})$*, with equality if and only if* $c_i = 1/n$ *for all* $1 \leq i \leq n$*.*

This result demonstrates that knowledge of the exact competence of agents is not required for improving decisions with respect to the straight average baseline. Rather, as long as the competence is ranked in the right order, the differentially weighted estimate will outperform the straight average.

The following result extends this finding to a case where the benefits of differential weighting are harder to anticipate: we allow the c_i to lie in the entire $[1/n, c_i^*]$ interval, allowing for cases where the ranking of the group members is not represented correctly. One might conjecture that this phenomenon adversely affects performance, but this is not the case:

Theorem 5 (Second Baseline Result) *Let* $c_1 \ldots c_n \in [0, 1]$ *be such that* $\sum_{i=1}^{n} c_i = 1$*. In addition, let* $c_i \in [\frac{1}{n}; c_i^*]$ *(respectively* $c_i \in [c_i^*; \frac{1}{n}]$*) hold for all* $1 \leq i \leq n$*. Then the differentially weighted estimator* $\hat{\mu} := \sum_{i=1}^{n} c_i X_i$ *outperforms the straight average. That is,* $\mathrm{MSE}(\hat{\mu}) \leq \mathrm{MSE}(\bar{\mu})$*, with equality if and only if* $c_i = 1/n$ *for all* $1 \leq i \leq n$*.*

In other words, as long as the relative weights lie in between the equal weights and the optimal weights, the accuracy of the group judgment is increased. Even a fallacious competence ranking need not be harmful: the resulting estimate will still be better than straight averaging. Briefly, as long as there is a positive correlation between degrees of expertise and impact on the group judgment, the group does well to weigh the estimates differentially.

The litmus test for Klein and Sprenger's results are cases where some of their idealizing assumptions fail, e.g., independence or unbiasedness. For example, training, experience, risk attitude or personality structure may bias the agents' estimates into a certain direction. In assessing the impact of industrial development on a natural habitat, an environmentalist will usually come up with an estimate that significantly differs from the estimate submitted by an employee of a corporation that intends to exploit the habitat—even if both are intellectually honest and share the same information. In these circumstances, the agents should not be modeled as unbiased statistical estimators, but as estimators whose mean value is different from μ. However, as long as the differentially weighted bias is smaller or equal than the average bias, the baseline results remain valid and differential weighting still outperforms straight averaging (Section 3 in Klein and Sprenger 2015).

Consider now the case where agents are not independent, but where their opinions are correlated with each other, e.g., because they draw from similar information sources (e.g., Goldman 2001). This may happen because they use similar research methods or because they share information with each other. For this case, Klein and Sprenger show the following result:

Theorem 6 *Let $X_1, \ldots, X_n$ be unbiased estimators, that is, $\mathbb{E}[X_i] = \mu = 0$, and let the c_i satisfy the conditions of one of the baseline results, with $\hat{\mu}$ defined as before. Let $I \subseteq \{1, \ldots, n\}$ be a subset of the group members with the property*

$$\forall i, j \neq k \in I : c_i \geq c_j \;\Rightarrow\; \mathbb{E}[X_j X_k] \geq \mathbb{E}[X_i X_k] \geq 0. \tag{6}$$

(i) **Correlation vs. Expertise** *If $I = \{1, \ldots, n\}$, then weighted averaging outperforms straight averaging, that is, $\mathrm{MSE}(\hat{\mu}) \leq \mathrm{MSE}(\bar{\mu})$.*

(ii) **Correlated Subgroup** *Assume that $\mathbb{E}\left[X_i X_j\right] = 0$ if $i \in I$ and $j \notin I$, and that*

$$\frac{1}{|I|} \sum_{i \in I} c_i \leq \frac{1}{n} \sum_{i=1}^{n} c_i. \tag{7}$$

Then weighted averaging still outperforms straight averaging, that is, $\mathrm{MSE}(\hat{\mu}) \leq \mathrm{MSE}(\bar{\mu})$.

To fully understand this theorem, we have to clarify the meaning of condition (6). Basically, it says that in group I, experts are less correlated with other (sub)group members than non-experts.

Once we have understood this condition, the rest is straightforward. Part (i) states that if I equals the entire group, then differential weighting has an edge over averaging. That is, the benefits of expertise recognition are not offset by the perturbations that mutual dependencies may introduce. Arguably, the generality of the result is surprising since condition (6) is quite weak. Part (ii) states that differential weighting is also superior whenever there is no correlation with the rest of the group, and as long as the average competence in the subgroup is lower than the overall average competence (see equation (7)).

It is a popular opinion (e.g., Surowiecki 2004) that correlation of individual judgments is one of the greatest dangers for relying on experts in a group. To some extent, this opinion is reflected by the above theorem. However, expertise-informed group judgments may still be superior to straight averaging, as demonstrated by Theorem 6. Thus, the interplay of correlation and expertise is subtle and cannot be generalized easily.

Summing up, taking into account relative accuracy positively affects the epistemic performance of groups even if the ranking of experts is only partially reliable (Theorem 4 and 5). The result remains stable over several representative extensions of the model, such as various forms of bias, violations of independence, and over- and underconfident agents (e.g., Theorem 6). In particular, differential weighting is superior (i) if experts are, on average, less biased; (ii) if all agents share the same sort of bias; (iii) if experts are less correlated with the rest of the group than other group members. These properties may be surprising and demonstrate the stability and robustness of expertise-informed judgments, implying that the benefits of recognizing experts may offset the practical problems linked with that process. The parsimony of this model and the independence of specific distributional assumptions suggest that these qualitative phenomena are likely to occur in reality, too.

6 Expert Judgment Literature

While the preceding sections dealt with various formal models for describing opinion aggregation and group judgments, this section surveys a limited number of empirical results regarding the elicitation and practical use of individual expertise. In particular, we present empirical justifications for both egalitarian and differential strategies in opinion aggregation.

6.1 Justification of Egalitarian Approaches

The empirical literature on expert judgment reflects the divisions and problems of the formal literature. On the one hand, scholars have supported equal weighting and independent forecasts on grounds that giving more weight to a method or another is unjustified unless the results are already known. Let us take an example: Imagine that an economist and a political scientist are forecasting the next 50 years of growth-rates for China. The two scientists may not only use different prediction strategies, but also focus on different sources of evidence: The economist may be inclined to use model-projections from past time-series, while the political scientist may rely on his personal rules-of-thumb, and "eyeball" her estimates based on intuitions about likely historical trends of nations with an economic history similar to today's China. Who is to be weighted more?

Which of the two strategies is the most successful one cannot be known a-priori. If the world stays more or less the same, we may expect model-projections to be more accurate than eyeballing. But time-series projections

suffer from known "broken-leg problems" (Bishop and Trout 2005, 45-53); that is, they ignore data that is particularly disruptive of otherwise smooth time-series. A political economist might be inclined to take China's aging problem more seriously than someone who looked only at index numbers; for instance, as a warning of the fact that China may well hit a growth-rate plateau in the next few decades.

Typically, in the selection of expertise we can nonetheless rely on the fact that experts are already selected with respect to their past performance. So the sociologist might be preferred over the economist, or vice-versa, on grounds that she has forecast past geopolitical events more successfully. We could, that is, give differential weights on the basis of past performance. This strategy too has its own flaws. In order for past performance to be meaningful—my judgment is as likely to be correct in the future as it has been in the past—the class of forecasts we are considering must be the relevant one. But it is often difficult to know which class of forecasts is the relevant one for the kind of problem we are considering. Kitcher offers a good example of the problem of relevance: In a thorough reconstruction of the disagreements over climate change, Oreskes and Conway (2010) show how several climate change skeptics had in fact obtained their status of experts in fields other than those relevant to climate sciences (see Kitcher 2010).

Consider the following simplified illustration. Let us assume that the political economist, in the example above, is a very reliable long-term forecaster of geopolitical events—she has successfully forecasted the geopolitical situation of several countries many years in advance. The economist, on the other hand, can very reliably predict the trend of key economic indicators (GDP, inflation, public debt, etc.). On the one hand, when purely economic factors are more likely to affect China's GDP growth in the next 50 years—that is, assuming a relative stability of the geopolitical system—then the economist has probably a better track record than the political scientist. On the other hand, if we think that geopolitical events are more likely to shape the future of Chinese growth, then the political scientist will likely have a better shot at the correct forecast. What the example highlights is that the two track records of individual experts are not always comparable—see Reiss (2008, 38-41) and Martini (2014) for a more detailed discussion.

In sum, selecting experts on grounds of performance indicators, credentials, etc. can be like comparing apples and oranges. There are often no one-dimensional meters of comparison with which to assess expertise.

Worse still, a scale can be handpicked by a relevant interest group in order to favor a preferred (biased) outcome as the *right* outcome. For the foregoing reasons, whenever we do not have a clear picture of the problem, of the kind of expertise that is relevant to it, etc., then we have an argument for egalitarian aggregation mechanisms, that is: equal-weighted averaging. Equal-weighted averaging is discussed in extensively in Armstrong (2001) and typically used in the *Delphi Method* for aggregating opinions (see Dalkey 2002). In his *Principles of Forecasting* Armstrong recommends using "equal weights unless you have strong evidence to support unequal weighting of forecasts" (2001, 422). He refers to Clemen, who "conducted a comprehensive review of the evidence and found equal weighting to be accurate for many types of forecasting" (see Clemen 1989). However, Armstrong notices, "the studies that [Clemen] examined did not use domain knowledge." (2001, 422) It is on the caveat of 'domain knowledge' that differential aggregation models can, under certain conditions, be preferable to egalitarian ones, as the next section will illustrate.

6.2 Justification of Differential Approaches

In some cases using differential weighting is justified by the fact that some of the members in the group possess more knowledge than others, and are therefore more likely to give accurate judgments. This typically happens in contexts where *domain knowledge* is involved. Domain knowledge is knowledge that is typical of a specific subject, field of research, etc., and is therefore most likely in possession of those with advanced training or experience in that field. To give an example, forecasting the trend of an economic indicator by means of an econometric model involves domain specific knowledge of statistics and econometric modeling; unlike the case of forecasting general geopolitical trends, which involves domain-general, rather than domain-specific, knowledge. But let us take a different case. Cooke (1991, 159ff.) reports on an experiment conducted at a Dutch training facility for operators of large technological systems. The experiment involved highly trained professionals, and aimed at testing whether experience is correlated with calibration. Subjects were tested over general knowledge questions and domain-specific questions. Following are some examples of the two types[2].

- General knowledge questions

[2]The examples are taken from Cooke (1991) and from Alpert and Raiffa (1982), the latter being the source of some of Cooke's own experimental questionnaires.

  - What was the total egg production in millions in the U.S. in 1965?
  - What is the total number of students currently enrolled in the Doctoral Program at the Harvard Business School?

- Domain knowledge questions
  - What is the maximal efficiency of the Tyne RM1A gas turbine?
  - What is the maximum admissible intake temperature for gas in the Olympus power turbine?

The distinction between general knowledge and domain knowledge is not a clear-cut one, leaving much room for debate in places where the two might overlap considerably. Nonetheless, we can safely say that general knowledge questions can usually be answered by reasoning on items of information that are widely shared. When trying to answer the question of how many students are currently enrolled at Harvard's MBA, we might be helped by reasoning over the following questions: How many students are typically enrolled in a university? How many in a typical American university? How large the Harvard Business School can be with respect to a typical business school? What is the minimum size of a business school to operate efficiently? So reasoning over related questions is likely to take us close to a reasonable range—i.e.,the Harvard Business School has between 1000 and 5000 students.

Unlike general knowledge, domain knowledge involves, in the first place, a terminology that is only mastered with specific education or training. The average educated person probably has very little information on how to even express the efficiency of a turbine, viz. 'what are the units?'. It is even less likely that we could find anyone who is not versed in the field of engineering, or more specifically aerospace engineering, who possesses the relevant information for estimating the admissible intake temperature of a turbine, or the efficacy of a specific model of turbine like 'Tyne RM1A' (cf. above).

Domain knowledge problems are such that we can often make a distinction between "experts" and "laymen", and for which we can hope to find sensible differential weights to be used in aggregation. The two important question then remain, from an empirical perspective, *when* to use differential aggregation, as opposed to equal weights, and, if used, how to find appropriate weights. The formal literature can help us find aggregation rules that respect a number of formal desiderata. But empirical assessment is just as important in discovering the efficiency of different weighting schemas.

It should be noted here that there are two general strategies for selecting weights: *ex-ante* weighting and *ex-post* weighting (see Ashton and Ashton 1985). The former strategy looks at reasons for choosing a certain weighting scheme based on evidence that is anterior to the results of the weighting. For instance, a manipulable weighting function that excludes (or reduces the importance of) some experts based on irrelevant evidence may be discarded a priori. But choosing an ex-ante weighting scheme may not be enough: we may want to calibrate the weights by an empirical (i.e., ex-post) assessment of the weighting scheme. An influential a posteriori strategy for selecting weights is given in Cooke (1991): weights are found by averaging an expert's performance on a number of seed questions in his or her own field of expertise, and relative to the field of expertise that is deemed relevant to the problem at hand. Seed questions are selected among knowledge items that are relevant for the problem in which the facilitator is interested, but whose answers are not known in advance by the agent that is being assessed. A seed problem could be, for example, "estimate the probability of failure of the Tyne RM1A gas turbine under a certain level of mechanical stress".

In Cooke's methodology, weights relative to expert X are determined on the basis of expert X's performance on a "quiz" that the facilitator thinks is relevant for the problem she is interested in. An engineer who is asked to assess the risks related to a nuclear power plant will be assessed (and weighted) on the basis of seed questions related, for instance, to components of nuclear power plants, the materials used in the contraction of power plants, etc. There is clearly an extrapolation problem here: The seed questions, it is assumed, are indicative of expertise in the problem of interest. This is clearly something that needs to be evaluated case-by-case, when formulating seed questions, and there seems to be no a-priori method for deciding which seed questions are valuable and which ones are not. While it is illustrative of how we can asses a weighting schema *ex-post*, Cooke's method is limited in that it can only be applied in cases where the problem is defined at a very detailed level, and where relevant seed questions are available. The burden of proof is therefore on those who want to use differential weighting to prove their case (see also Armstrong 2001).

7 Conclusions

This article has given a survey over the problem of combining individual judgments into group judgments, with a distinct focus on individual expertise and differential weighting.

Our survey revealed that there are different conceptions of group rationality in opinion aggregation, and that their appropriate use depends on the context. For example, in democratic decision-making, egalitarian weighting schemes are usually compulsory. Whereas in scientific contexts, considerations of expertise or relevant experience (e.g., in medical diagnosis) may be more forceful and motivate the use of a differential weighting scheme. The epistemic benefits and drawbacks of such schemes have been investigated theoretically in Section 5 and empirically in Section 6.

It also transpired that there is not necessarily an all-encompassing account of group rationality. The impossibility results in Section 2 and 3 show that for a logically interconnected agenda of propositions, there may be no opinion aggregation procedure that satisfies a set of plausible and intuitive constraints. These results hold for the aggregation of binary judgments as well as for the aggregation of graded attitudes, e.g., degrees of belief. This was actually one of the reasons why we have investigated the problem of aggregating a single numerical estimate in Section 4 and 5. While Section 4 investigated the rationality of belief revision and reaching a consensus, Section 5 gave a general epistemic analysis of differential weighting schemes, albeit in a simple statistical model.

It is a typical feature of the literature on opinion aggregation and expertise that there is a variety of approaches spread over different disciplines and research methods: the same problem may be tackled from the point of view of social choice theory, (formal) epistemology, mathematics and statistics, economics, experimental psychology, management science, and risk studies. Therefore our survey necessarily remains incomplete. However, we hope to have given the reader a taste of the diversity of the approaches to modeling individual expertise, as well as of the enormously complex interplay between formal models of group judgments and empirical studies. In particular, we hope to have created a better understanding of the differences between egalitarian and differential models of opinion aggregation and group decision-making, and of the philosophical rationales behind them.

Acknowledgements

The authors thank the editors of this book and two anonymous referees for their useful feedback. Jan Sprenger wishes to thank the Netherlands Organisation for Scientific Research (NWO) for support of his research through Vidi grant #276-20-023.

References

Alpert, Marc, and Howard Raiffa (1982): "A progress report on the training of probability assessors.", in: Amos Tversky, Paul Slovic and Daniel Kahneman (eds.) (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Armstrong, J. Scott (2001): "Combining Forecasts", in: J. Scott Armstrong (ed.) (eds.), *Principles of Forecasting: A Handbook For Researchers and Practitioners*. Norwell, MA: Kluwer Academic Publishers.

Ashton, Alison Hubbard, and Robert H. Ashton (1985): "Aggregating subjective forecasts: Some empirical results", *Management Science* 31(12): 1499-1508.

Bates, J.M., and C.W.J. Granger (1969): "The combination of forecasts", *Operational Research Quarterly* 20: 451–468.

Baumann, Michael R., and Bryan L. Bonner (2004): "The effects of variability and expectations on utilization of member expertise and group performance", *Organizational Behavior and Human Decision Processes* 93: 89–101.

Bishop, Michael A., and John D. Trout (2005): *Epistemology and the psychology of human judgment*. Oxford: Oxford University Press.

Bonner, Bryan L. (2004): "Expertise in Group Problem Solving: Recognition, Social Combination, and Performance", *Group Dynamics: Theory, Research, and Practice* 8: 277–290.

Bonner, Bryan L., Michael R. Baumann, and Reeshad S. Dalal (2002): "The effects of member expertise on group decision-making and performance", *Organizational Behavior and Human Decision Processes* 88: 719–736.

Bonner, Bryan L., Sheli D. Silito and Michael R. Baumann (2007): "Collective estimation: Accuracy, expertise and extroversion as sources of intra-group influence", *Organizational Behavior and Human Decision Processes* 103: 121–133.

Bovens, Luc, and Wlodek Rabinowicz (2006): "Democratic answers to complex questions: an epistemic perspective", *Synthese* 150: 131–153.

Bradley, Richard, Franz Dietrich, and Christian List (2014): "Aggregating Causal Judgments", *Philosophy of Science* 81: 491–515.

Brössel, Peter, and Anna-Maria A. Eder (2014): "How to resolve doxastic disagreement", *Synthese* 191: 2359–2381.

Christensen, David (2009): "Disagreement as Evidence: The Epistemology of Controversy", *Philosophy Compass* 4: 756–767.

Clemen, Robert T. (1989): "Combining forecasts: a review and annotated bibliography", *International Journal of Forecasting* 5: 559–583.

Cooke, Roger M. (1991): *Experts in uncertainty*. Oxford: Oxford University Press.

Dalkey, Norman C. (2002): "A Delphi Study of Factors Affecting the Quality of Life", in: Harold A. Linstone and Murray Turoff (eds.), *The Delphi Method. Techniques and Applications*, Boston. Addison-Wesley.

Davis, James H. (1973): "Group decision and social interaction: A theory of social decision schemes", *Psychological Review* 80: 97–125.

Davis, James H. (1996): "Group decision making and quantitative judgments: A consensus model", in: E. Witte and J.H. Davis (eds.), *Understanding group behavior: Consensual action by small group*, 35–59. Mahwah, NJ: Erlbaum.

DeGroot, Morris (1974): "Reaching a Consensus", *Journal of the American Statistical Association* 69: 118–121.

Dietrich, Franz, and Christian List (2008): "Judgment aggregation without full rationality", *Social Choice and Welfare* 31: 15–39.

Einhorn, Hillel J., Robin M. Hogarth, and Eric Klempner (1977): "Quality of Group Judgment", *Psychological Bulletin* 84: 158–172.

Elga, Adam (2007): "Reflection and Disagreement", *Noûs* 41: 478–502.

Faust, David (1984): *The Limits of Scientific Reasoning*. Minneapolis: University of Minnesota Press.

French, John R.P. Jr. (1956): "A Formal Theory of Social Power", *Psychological Review* 63: 181–194.

Geanakopoulos, John D., and Heraklis M. Polemarchakis (1982): "We cannot disagree forever", *Journal of Economic Theory* 28: 192–200.

Genest, Christian, and James V. Zidek (1986): "Combining probability distributions: A critique and an annotated bibliography", *Statistical Science* 1: 114–135.

Goldman, Alvin (2001): "Experts: Which Ones Should You Trust?", *Philosophy and Phenomenological Research* 63: 85-110.

Grossi, Davide (2009): "Unifying preference and judgment aggregation", *AAMAS* 1: 217–224.

Hartmann, Stephan, Gabriella Pigozzi, and Jan Sprenger (2010): "Reliable Methods of Judgment Aggregation", *Journal for Logic and Computation* 20: 603–617.

Hartmann, Stephan, and Jan Sprenger (2012): "Judgment Aggregation and the Problem of Tracking the Truth", *Synthese* 187: 209–221.

Helmer-Hirschberg, Olaf (1967): "Analysis of the Future: The Delphi Method". Working paper of the RAND Corporation.

Hinsz, Verlin B. (1999): "Group Decision Making with Responses of a Quantitative Nature: The Theory of Social Decision Schemes for Quantities", *Organizational Behavior and Human Decision Processes* 80: 28–49.

Hogarth, Robin M. (1978): "A Note on Aggregating Opinions", *Organizational Behavior and Human Performance* 21: 40–46.

Kelly, Thomas (2005): "The Epistemic Significance of Disagreement", in: J. Hawthorne and T. Szabo (eds.), *Oxford Studies in Epistemology*, 167–196. Oxford: Oxford University Press.

Kitcher, Philip (2010): "The Climate Change Debates", *Science* 328: 1230-1234.

Klein, Dominik, and Jan Sprenger (2015): "Modeling individual expertise in group judgments", forthcoming in *Economics and Philosophy*.

Kuhn, Thomas S. (1962): *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuhn, Thomas S. (1977): *The Essential Tension*. Chicago: University of Chicago Press.

Laughlin, Patrick R., and Alan L. Ellis (1986): "Demonstrability and social combination processes on mathematical intellective tasks", *Journal of Experimental Social Psychology* 22: 177–189.

Laughlin, Patrick R., Erin C. Hatch, Jonathan S. Silver, and Lee Boh (2006): "Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Effects of Group Size", *Journal of Personality and Social Psychology* 90: 644–651.

Lehrer, Keith (1976): "When Rational Disagreement is Impossible", *Noûs* 10: 327–332.

Lehrer, Keith, and Carl Wagner (1981): *Rational Consensus in Science and Society*. Reidel: Dordrecht.

Lehrer, Keith, and Carl Wagner (1983): "Probability amalgamation and the independence issue: a reply to Laddaga", *Synthese* 55: 339–346.

Libby, Robert, Ken T. Trotman, and Ian Zimmer (1987): "Member variation, recognition of expertise and group performance", *Journal of Applied Psychology* 72: 81–87.

Lindley, Dennis V. (1983): "Reconcilation of probability distributions", *Operations Research* 31: 866–880.

Linstone, Harold A., and Murray Turoff (1975): *The Delphi Method: Techniques and Applications*. Reading/MA: Adison-Wesley.

Littlepage, Glenn E., Greg W. Schmidt, Eric W. Whisler, and Alan G. Frost (1995): "An input-process-output analysis of influence and performance in problem-solving groups", *Journal of Personality and Social Psychology* 69: 877–889.

List, Christian (2005): "Group knowledge and group rationality: a judgment aggregation perspective", *Episteme* 2: 25–38.

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

List, Christian, and Philip Pettit (2002): "Aggregating sets of judgments: An impossibility result", *Economics and Philosophy* 18: 89–110.

List, Christian and Clemens Puppe (2009): "Judgment aggregation: a survey", in: Paul Anand, Prasanta Pattanaik and Clemens Puppe (eds.), *The Handbook of Rational and Social Choice*, 457-482. Oxford: Oxford University Press.

Loewer, Barry, and Robert Laddaga (1985): "Destroying the Consensus", *Synthese* 62: 79–95.

Lorge, Irving, and Herbert Solomon (1959): "Individual performance and group performance in problem solving related to group size and previous exposure to the problem", *The Journal of Psychology* 48: 107–114.

Martini, Carlo (2014): "Experts in Science: A View From the Trenches", *Synthese* 191: 3–15.

Martini, Carlo, Jan Sprenger, and Mark Colyvan (2013): "Resolving Disagreement Through Mutual Respect", *Erkenntnis* 78: 881–898.

McConway, K. J. (1981): "Marginalization and linear opinion pools", *Journal of the American Statistical Association* 76: 410–414.

Nadeau, Richard, Cloutier, Edouard, and Guay, J.-H. (1993): "New Evidence About the Existence of a Bandwagon Effect in the Opinion Formation Process", *International Political Science Review* 14: 203–213.

Nitzan, Shmuel, and Jacob Paroush (1982): "Optimal decision rules in uncertain dichotomous choice situations", *International Economic Review* 23(2): 289–297.

Oreskes, Naomi, and Erik M. Conway (2010): *Merchants of doubt*. London: Bloomsbury Publishing.

Page, Scott E. (2007): *The Difference*. Princeton: Princeton University Press.

Reiss, Julian (2008): *Error in Economics: Towards a More Evidence-Based Methodology*. New York: Routledge.

Romeijn, Jan Willem, and Olivier Roy (2015): "All Agreed: Aumann Meets DeGroot". Unpublished manuscript.

Shapley, Lloyd, and Bernard Grofman (1984): "Optimizing group judgmental accuracy in the presence of interdependencies", *Public Choice* 43: 329–343.

Surowiecki, James (2004): *The Wisdom of the Crowds*. Harpswell: Anchor.

Thomas, Edwin J., and Clifton F. Fink (1961): "Models of group problem solving", *The Journal of Abnormal and Social Psychology* 63: 53–63.

Trotman, K.T., P.W. Yetton and I.R. Zimmer (1983): "Individual and group judgments of internal control systems", *The Journal of Accounting Research* 21: 286–292.

Trout, J.D. (2009): *The empathy gap: Building bridges to the good life and the good society*. New York: Viking/Penguin.

Wagner, Carl (1978): "Consensus through respect: a model of rational group decision-making", *Philosophical Studies* 34: 335–349.

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

Zajonc, Robert B., and William H. Smoke (1959): "Redundancy in task assignments and group performance", *Psychometrika* 24: 361-369.