J Philos Logic (2012) 41:983–999
DOI 10.1007/s10992-012-9227-5

Testimony as Evidence: More Problems for Linear Pooling

Katie Steele

Received: 15 September 2011 / Accepted: 15 January 2012 / Published online: 4 May 2012

Abstract This paper considers a special case of belief updating—when an agent learns testimonial data, or in other words, the beliefs of others on some issue. The interest in this case is twofold: (1) the linear averaging method for updating on testimony is somewhat popular in epistemology circles, and it is important to assess its normative acceptability, and (2) this facilitates a more general investigation of what it means/requires for an updating method to have a suitable *Bayesian* representation (taken here as the normative standard). The paper initially defends linear averaging against Bayesian-compatibility concerns raised by Bradley (Soc Choice Welf 29:609–632, 2007), as well as problems associated with *multiple* testimony updates. The resolution of these issues, however, requires an extremely nuanced interpretation of the parameters of the linear averaging model—the so-called *weights of respect*. We go on to propose a role that the parameters of any 'shortcut' updating function should play, by way of minimal interpretation of these parameters. The class of updating functions that is consistent with this role, however, excludes linear averaging, at least in its standard form.

Keywords Testimony · Linear pooling · Bayesian belief change

1 Introduction

This paper has two goals. The first, more specific goal, is to investigate one popular method for updating belief in response to the testimony of others: the

K. Steele (✉)
London School of Economics and Political Science, London, UK
e-mail: k.steele@lse.ac.uk

linear averaging (or *linear pooling*) method. Testimony is here understood as another agent's (Bayesian) beliefs, whether the beliefs concern direct experience, or a more involved inference, for instance, that there is life elsewhere in the universe. The second goal of the paper is more general: to investigate what it means for a belief-updating method to be *viable, with respect to the Bayesian model,* which is taken here as the normative standard. The testimony case is useful for investigating this general issue, because testimony has been treated as a special case of learning in the literature, and has inspired alternative, not-obviously-Bayesian, belief-updating methods, such as linear pooling.

A few words on the relevant testimony literature are in order. One point of connection is the recent debate concerning how to resolve 'peer disagreement' that persists after all evidence has been shared (see, for instance, Feldman [7], Elga [6], Christensen [3] and Kelly [13]). There is also a more formal literature (the actual starting point for this paper) concerning models of learning that specifically handle input data in the form of others' probabilistic beliefs (see Lehrer and Wagner [14], French [9], Genest and Zidek [10], Clemen and Winkler [4]).[1] In this literature too, it is assumed that agents have, as far as possible, shared all relevant data—at the time in question, they are in *reflective equilibrium*, to use the terminology of Lehrer and Wagner [14]. We will follow suit and restrict attention to cases where agents have shared all background evidence, at least to the best of their knowledge. In fact, the situation can be envisaged as one where agents update on a portion of each others' *prior beliefs*. Of course, the implicit assumption is that the agents do not have the same prior beliefs over an entirely identical probability space, and/or the same interpretation of the evidence, otherwise there would be no persistent disagreement.[2] Despite these differences, it is plausible that agents may yet regard each other as epistemic peers, in the sense that they *learn* from aspects of each others' prior beliefs.[3]

One might be concerned about just what counts as testimony, and whether it is really distinct from other kinds of data. Surely there are multitudes of more and less explicit ways to learn of another's beliefs, such as observing their behaviour, or mere traces of their behaviour. For example, the fact that my friend has taken her umbrella tells me something about her belief that it will

[1]The cited models actually play a more ambitious role; they are intended to model the process of *consensus*, where a number of group members update on each others' beliefs and arrive at the same belief. Here we are concerned only with the proposed methods of updating, and not the conditions under which consensus is achieved within a group.

[2]Likewise, there can be no ultimate disagreement if agents have the same prior probability function and have common knowledge of each other's posteriors, according to Aumann's [1] theorem.

[3]This is a slightly more general interpretation of 'epistemic peer' than in the informal peer disagreement debate; in that debate, an epistemic peer is often taken to be someone whose beliefs one respects *equally* to one's own.

rain. Should all these be treated as special cases of learning testimony? Perhaps the best response to this question is: yes. Even if it is gleaned in a variety of ways, testimony can be similarly represented as a statement of the beliefs of others. That is, testimony is not 'raw' data, but is the result of an implicit initial inference, as shown below:

$$\text{event (e.g. communication)} \rightarrow \text{initial inference} \rightarrow \text{testimony}$$

The general testimony problem is as follows: the agent in question has a prior probabilistic belief function over a sigma algebra of events. The agent receives some testimonial data *at some point*, i.e. they learn the beliefs, regarding the partition $\mathbf{B}$, of n other agents, the respective belief functions being $P_i, i = 1 \ldots n$. (These others need not entertain the same full probability space as our agent of interest; they must merely have beliefs on $\mathbf{B}$.) By way of representation, we can construct a matrix $M_{\mathbf{B}}$ with columns corresponding to the events in $\mathbf{B}$ and rows corresponding to the belief functions of the expert peers. (We can refer to this matrix as the testimony profile.) The task is to specify the principal agent's posterior belief function. That is, what is an appropriate updating function $F_{\mathbf{B}}$ for determining our agent's new beliefs across $\mathbf{B}$, given their prior beliefs and the testimony profile $M_{\mathbf{B}}$? To state the problem semi-formally, the general problem is to find plausible candidates for:

$$F_{\mathbf{B}}\text{: prior, } M_{\mathbf{B}} \rightarrow \text{posterior}$$

As suggested earlier, our guiding aim in what follows is to examine one function $F_{\mathbf{B}}$ that has become popular in both the peer disagreement and formal consensus literature—the weighted linear average.[4] Not only has testimonial data been taken as a distinct form of data, it has also been considered worthy of special treatment vis-à-vis updating; hence the focus on linear pooling. Section 2 briefly comments on why this might be so.

The rest of the paper examines whether and how linear averaging may be compatible with Bayesian updating. Section 3 argues, against Bradley [2], that an instance of linear averaging *can* be considered prima facie compatible with an instance of Bayesian conditioning. The 'trick' is to carefully interpret the parameters of the averaging function, i.e., the *weights of respect*. Further issues arise, however, when we turn to *multiple* testimony updates on different issues. Section 4 argues that unless weights of respect are even more 'loaded', linear averaging does not adequately treat testimony as incremental evidence, such that commutativity of multiple testimony updates is assured. Sections 5 and 6 take stock of these interpretative issues regarding weights of respect;

[4]Weighted linear averages are prominent in the formal consensus modeling literature. (Geometric averages are also considered in this literature and will be acknowledged in later sections of the paper.) A crude kind of linear average is arguably the favoured solution in the peer disagreement debate—what is referred to as the 'equal weights' view.

we propose a role that the weights, or more generally, the parameters of an updating function, should play. A class of updating functions is identified that is consistent with this role, but the class includes only a modified version of linear averaging, and not linear averaging proper. Some further concerns about sequential testimony updates are acknowledged in Section 7, and the concluding Section 8 reflects on the prospects for linear pooling and its cousins.

2 Why not Bayesian Business as Usual?

Before turning to an examination of linear averaging for updating on testimony, it is helpful to consider why alternatives to Bayesian conditionalization might have been proposed in the first place. The aim of this section is merely to offer some initial motivation for departures from the standard Bayesian model.

The Bayesian (on at least one reading) holds that the pooling function $F_{\mathbf{B}}$ should accord with Bayes' formula:

$$\begin{aligned} P_0'(B_j) &= P_0(B_j | P_1(B_j) = p_1, P_2(B_j) = p_2, \ldots, P_n(B_j) = p_n) \\ &= \frac{P_0(P_1(B_j) = p_1, P_2(B_j) = p_2, \ldots, P_n(B_j) = p_n | B_j) \times P_0(B_j)}{P_0(P_1(B_j) = p_1, P_2(B_j) = p_2, \ldots, P_n(B_j) = p_n)} \\ &\quad \forall B_j \text{ in } \mathbf{B}. \end{aligned}$$

where $P_0(B_j)$ and $P_0'(B_j)$ are the agent's prior and posterior for event B_j and $P_i(B_j)$ is witness i's probability for event B_j (at the time in question).

The 'problem' with this function, or the reason some might consider it not sufficiently user-friendly, is that the testimony of others is not combined directly with the agent's own probabilistic beliefs; belief change is governed, rather, by the relevant likelihoods

$$P_0(P_1(B_j) = p_1, P_2(B_j) = p_2, \ldots, P_n(B_j) = p_n | B_j)$$

which represent the agent's belief that their peers would have the beliefs specified, conditional on the truth of each event B_j in $\mathbf{B}$. (The likelihoods conditional on the falsity of each B_j also play a role, of course.) We do not here deny that Bayesian conditioning is the most accurate way to represent rational belief change in response to testimony, in a sufficiently detailed model; indeed the Bayesian model is treated as the normative standard in this paper. The point is just that the Bayesian model may be somewhat cumbersome (with respect to the number of propositions that must be modeled) and also awkward to use. Indeed, the Bayesian expression above treats testimony just like any other type of evidence—an event that is merely indicative of the truth/falsity of the events B_j under consideration.

Averaging methods may have become popular 'shortcuts' for updating on testimony, precisely because, in contrast to the above, the probabilistic beliefs

of others play a direct role in these functions. We see this by considering the formal statement of the weighted linear average:

$$P'_0(B_j) = w_0 \times P_0(B_j) + w_1 \times P_1(B_j) + \ldots + w_n \times P_n(B_j)$$

where $w_0, \ldots, w_n$ are interpreted as the 'weights of respect' assigned to all agents involved, and are non-negative and summing to one.

The posterior belief on **B** for the agent in question is a linear 'pool' of the actual beliefs of all agents involved. Instead of treating other agents' beliefs like a litmus test for determining the truth of the events B_j, the agent takes these beliefs on board directly; the agent mixes these beliefs directly with their own. Weighted linear averaging allows this to be done in such a way that the beliefs of those the agent most respects have the most influence, or are most dominant in the mix. This is, *prima facie*, a more natural or user-friendly way to respond to the beliefs of others.

The popular defence for using linear averaging, in particular, to serve as the shortcut function for updating on testimony ($F_{\mathbf{B}}$), is given mathematically in Wagner [16]. Lehrer and Wagner [14] mount the same defence more explicitly in the context of the updating problem as opposed to the group aggregation problem. In short, the linear average is the only function to satisfy the *Independence of Irrelevant Alternatives* condition, (*IA*). The appeal to IA amounts to a *multi-profile* justification: if we consider all possible combinations of probability functions for the agents involved, IA states that where the vector of probabilities for a single event B_j are equivalent, the posterior belief for B_j should be equivalent. In other words, the posterior for a single event B_j depends just on the probabilities all agents assign to B_j, and it does not matter what are the complete belief functions of these agents. Wagner [16] proves that if IA alone is stipulated (under *universal domain*), then $F_{\mathbf{B}}$ must be a weighted linear average, with some error term. The function is restricted to positive weights and zero error (as per the expression above) if the further condition of *Zero Preservation*, (*ZP*), is stipulated; ZP states that if all agents assign probability zero to some event B_j, then the posterior for B_j should also be zero.

One might regard the above two-step explanation a bit quick, i.e. that there is sufficient motivation for a special method for updating on testimony (which is moreover a distinct type of information), and that this method should be a weighted linear average, as it must satisfy IA and ZP.[5] Even if one sees

[5]Indeed, fans of the weighted geometric average would reject the latter desiderata—IA. The geometric averaging method has the following form:

$$P'_0(B_j) = \text{ normalise } [P_0(B_j)^{w_0} \times P_1(B_j)^{w_1} \times \ldots \times P_n(B_j)^{w_n}]$$

where $w_0, \ldots w_n$ are non-negative and summing to one.

There is some debate about the comparative merits of linear and geometric averaging; see Genest and Zidek [10], Clemen and Winkler [4], Shogenji [15], Jehle and Fitelson [12]. An aspect of the debate is noted in the next section.

little positive motivation for linear averaging, however, the fact remains that the method has a significant presence in the literature, and it is important to consider whether this updating approach is at least rationally *permissible*. To this end, the next sections examine whether linear averaging is compatible with the Bayesian model.

3 Single Testimony Updates: An Initial Compatibility Challenge

The task of determining whether linear averaging is compatible with Bayesian conditionalization mostly involves working out what it *means* for the two updating methods to be compatible. Our initial examination of compatibility in this section is confined to a single testimony update in isolation.

Linear averaging is prima facie a special case of Bayesian conditionalization. To begin with, not all Bayesian updates on testimony can be expressed as linear averages over the relevant partition. For example, there is some Bayesian model, yet no linear averaging model, that permits an agent the following update:

$$\text{prior } [\,1/6\ 1/3\ 1/2\,] \text{ plus testimony } [\,1/2\ 1/6\ 1/3\,] \longrightarrow [\,1/3\ 1/3\ 1/3\,]$$

On the other hand, any singular instance of linear averaging can be represented by *some* Bayesian model. It is a matter of equating:

$$P'_0(B_j) = w_0 \times P_0(B_j) + \ldots + w_n \times P_n(B_j) \quad \text{and}$$
$$P'_0(B_j) = P_0(B_j|M_{\mathbf{B}}(j)) = \frac{P_0(M_{\mathbf{B}}(j)|B_j) \times P_0(B_j)}{P_0(M_{\mathbf{B}}(j))}$$

If all the terms in the linear average expression are fixed, then we have the following constraint on the Bayesian likelihood ratio (where k is a function of the averaging weights and the priors of all concerned, including the principal agent):

$$P_0(M_{\mathbf{B}}(j)|B_j)/P_0(M_{\mathbf{B}}(j)|\neg B_j) = k$$

Since there are many ways for a ratio to equal some constant k, there is in fact a one-to-many relationship between averaging updates on some specified testimony and equivalent Bayesian updates on the same testimony. So the averaging representation of a single testimony update under-determines the corresponding Bayesian representation. As far as compatibility goes, so far so good—the important thing to note here is that there is *some* Bayesian model or class of models that can represent an isolated linear-averaging update.

So much for equating linear averaging and Bayesian models in the abstract. The further question is whether the models are compatible vis-à-vis the real world. Do the 'right' instances of these models yield equivalent posteriors, so to speak? Bradley [2] effectively argues that there is an obvious way in which the interpretations of the models do not line up. This section defends averaging against Bradley's and similar criticisms, by showing that the problem can be dodged by attending carefully to the universal domain condition that grounds the averaging method. There remain further problems for linear averaging, but these must wait for the next section.

A number of authors question how 'weights of respect' in linear averaging should be interpreted and ascertained (e.g. French [9]), but Bradley [2] expresses a more fundamental worry about these weights: If we appeal to the multi-profile justification of linear averaging given in the last section, then the respect weights are constant for group members regardless of the nature/origins of the beliefs they express, and thus the method ignores real-world distinctions that are both salient and important in the Bayesian setting.

In particular, Bradley argues that averaging treats independent agents identically to dependent agents, while the Bayesian treats them differently. Let us rehearse the argument. Following Bradley, we will keep things simple and assume there are just two agents giving testimony to the principal agent. The latter learns the probabilities for these two agents across partition **B**. According to the Bayesian model, our agent's new probability for one event in **B**, call it b, given this new information, is:

$$\begin{aligned} &P_0(b \,|\, P_1(b) = p_1,\, P_2(b) = p_2) \\ &\quad = \frac{P_0(P_1(b) = p_1,\, P_2(b) = p_2|b) \times P_0(b)}{P_0(P_1(b) = p_1,\, P_2(b) = p_2)} \end{aligned}$$

If the probability functions for the two consulted experts are independent given b, and are also unconditionally independent, the above equals:

$$\frac{P_0(P_1(b) = p_1|b) \times P_0(P_2(b) = p_2|b) \times P_0(b)}{P_0(P_1(b) = p_1) \times P_0(P_2(b) = p_2)}$$

At the other extreme, if the probability functions for the two experts are perfectly correlated, the expression equals:

$$\frac{P_0(P_1(b) = p_1|b) \times P_0(b)}{P_0(P_1(b) = p_1)}$$

Bradley notes that these two expressions will only be equal if

$$\frac{P_0(P_2(b) = p_2|b)}{P_0(P_2(b) = p_2)} = 1$$

which is to say that the principal agent believes one of the agents' beliefs to be independent not only of the other agent, but also independent of the truth. In that case, by any reasonable interpretation of weights of respect, this agent should be given a weight of zero in the linear averaging function.

We see from the above that if the domain of an averaging method is any probability profile for a group, including cases where the experts consulted have independent beliefs as well as cases where their beliefs are thought to be dependent in some way, then the linear averaging and Bayesian representations of testimony in the real world will only be compatible in special circumstances—when the set of respect weights mirrors the situation where the beliefs of all consulted experts except one are independent of the truth (the trivial case as far as independence versus dependence amongst experts is concerned). This sort of restriction defeats the purpose of providing a model that allows an agent to update on the beliefs of *a number* of other agents who they regard as epistemic peers.

A response can be made to the criticism above, that does not involve sacrificing the multi-profile justification of linear averaging. The trick is to carefully specify the domain over which the IA condition must hold, and consequently the domain over which the weights of respect are constant. Consider the following situation: Our agent may consult Group 1, constituted by experts whose beliefs about **B** are independent, or else our agent may consult Group 2, constituted by experts whose beliefs have some pattern of dependency. The IA/universal-domain condition effectively requires the same respect weights be assigned to members of Group 1 (or 2), whatever the members' belief functions happen to be. It does not, however, require matching respect weights across two different groups who express the same set of probability functions, unless the testimony profiles associated with these groups are not distinguished, but are rather part of the same domain. And this need not be the case.

This point about the domain of a particular linear averaging function applies more broadly than the case of independent versus (partially) dependent peer groups. One might also be worried that the same group of experts may have greatly different expertise with respect to different issues, and yet constant weights of respect will not reflect this. For instance, we would not want to assign the same respect weightings for a certain group of peers regardless of whether we were asking them about average rainfall for the next wet season or whether unemployment will drop. Again, the mistake here is to think that belief profiles concerning different issues/propositions are part of the same domain over which the justifying condition for linear averaging, IA, applies. It must simply be stipulated that IA applies only to sets of beliefs concerning the *same* event space. That is, if the partition **B** is being assessed, the universal domain spans all possible (prior, testimony profile) pairs for the group that are constituted by probability functions on **B**. If, on the other hand, a different partition is in question, say **C**, then the universal domain would span

a different set of (prior, testimony profile) pairs, this time probability functions on **C**.[6]

Given the points just made, a slight change in the notation used above is in order. Recall the general representation of the testimony problem given in the first section:

$$F_{\mathbf{B}} : \text{ prior}, M_{\mathbf{B}} \rightarrow \text{ posterior}$$

The use of the **B** index reflects the point above that the matrix of probability functions in question pertains to a specific issue or partition—the **B** issue. The updating function (in particular the respect weights) are specific to that issue, hence $F_{\mathbf{B}}$. But we might want to make explicit the first point as well, that the matrix is specific to a particular context—a group of peers at a particular point with a believed pattern of dependency in their beliefs. This rich context might be represented by the further index $G \in \mathbf{G}$. So the testimony received by an agent is represented $M_{\mathbf{B},G}$, and likewise, the updating function that we seek is $F_{\mathbf{B},G}$. Of course, all this indexing highlights how removed testimony is from raw experience in our model; the evidential statement $M_{\mathbf{B},G}$ is laden with inferences about the 'group context' that are not explicitly modeled. To put the point another way: it is not clear how the group context is translated into weights of respect. We put these issues aside for the moment, but will return to them later in Section 5.

4 Rich Event Spaces and Bayesian Compatibility

The reference above to different issues or partitions, say, **B** and **C**, raises the further question of how linear averaging is supposed to work in a rich algebraic setting. Unless this is clarified, averaging belief-update methods are at best incomplete. Arguably the most natural solution is that averaging be combined with Jeffrey-updating, such that, in response to testimonial data regarding **B**,

[6]Note that we might require that the same weights of respect apply to partitions **B** and **B**′, where the latter is a refinement of the former. In this way, the posterior probabilities for the events in **B** will be identical in each case, whether updating is performed on this coarse partition directly, or on the refined partition. Indeed, linear averaging is championed (over geometric averaging) for having this *marginalization* property. Geometric averaging, by contrast, satisfies a different invariance property; it has what Wagner refers to as the *Bayesianity* property. That is, if a geometric update with some set of respect weights is performed on a complex space $\mathbf{B} \times \mathbf{E}$, and then some 'ordinary' evidence E_i is learnt (via conditionalization), the posterior distribution is the same as if the expert peers all learnt E_i first (via conditionalization), and then the principal agent did a geometric update on $\mathbf{B} \times \mathbf{E}$, with weights as before. Linear averaging does not have the Bayesianity property, stated as such, but this may be mitigated if the weights assigned to expert peers may change, depending on whether or not they have already learnt E_i. See, for instance, Genest and Zidek [10], Clemen and Winkler [4], Shogenji [15] and Jehle and Fitelson [12] for discussion of these two properties of testimony updating functions.

probabilities are updated across this partition in line with linear averaging, and then the probabilities of all other propositions are subsequently updated so that probabilities conditional on the individual events of **B** remain constant or 'rigid'.

It is best to illustrate with an example. Consider a simple setting where our agent has the following prior probability function $P_{0,\mathbf{D}}$ over the event space $\mathbf{D} = \mathbf{B} \times \mathbf{C}$:

	B	$\neg B$
C	0.1	0.2
$\neg C$	0.3	0.4

The agent meets an expert on the **B** partition, who has $P_{1,\mathbf{B}} = [0.8, 0.2]$. Linear pooling with, for example, weight 0.5 to this expert, gives $[0.6, 0.4]$ over the **B** partition. Clearly, however, this is not a complete specification of the agent's posterior probability function. This is where Jeffrey conditionalization enters. Accordingly, the agent's new probabilities over the entire space are:

	B	$\neg B$
C	0.15	$0.1\dot{3}$
$\neg C$	0.45	$0.2\dot{6}$

The two-step procedure—averaging then Jeffrey-conditionalization—can thus be regarded a fully comprehensive belief-update rule!

We noted in the previous section that many Bayesian models yield the same posterior distribution across **B** as some averaging update on **B** in response to testimony. The Jeffrey-conditionalisation step can be considered a further constraint on, or specification of, averaging updates, such that the class of consistent Bayesian models is smaller. Alternatively, one might conceive of the Jeffrey move as positioning linear-averaging as an *extra*-Bayesian process. Learning testimony is no longer likened to gaining knowledge of a single proposition E, as per strict Bayesian conditionalization. Rather, learning testimony results in a change (via averaging) in the probability distribution across some partition; the testimonial evidence itself is not explicitly modeled as a proposition.

The analysis of rich event spaces does not, however, end here. The well-known puzzles with Jeffrey-conditionalization—that it is not generally commutative with respect to changes in probabilities on different partitions—prompt further investigation of the proper treatment of testimony. Our earlier example can be extended to illustrate the non-commutativity property. Recall that our agent has already updated her beliefs with respect to **B** in response to the beliefs of an expert peer on **B**. After this encounter, assume that the agent meets a (different) expert on the **C** partition who has $P_{2,\mathbf{C}} = [0.6, 0.4]$.

The obvious strategy is to apply the linear-plus-Jeffrey method a second time around. Linear pooling, again with the assumption of weight 0.5 for

the **C**-expert, gives [0.44167, 0.55833] over the **C** partition. Applying Jeffrey conditionalization, we get the following posterior:

	B	$\neg B$
C	0.23383	0.20784
$\neg C$	0.35058	0.20775

But what if the experts had made their reports in reverse order? *If we assume the same weightings should be applied to the experts*, we would now get the following transition of probability functions:

	B	$\neg B$			B	$\neg B$			B	$\neg B$
C	0.1	0.2	$\rightarrow$	C	0.15	0.30	$\rightarrow$	C	0.23056	0.19884
$\neg C$	0.3	0.4		$\neg C$	0.23571	0.31429		$\neg C$	0.36230	0.20830

There is not a large difference in the posterior matrices for this particular example. Nonetheless, the averaging-plus-Jeffrey update procedure, as conceived above with fixed weightings, yields posteriors that are sensitive to the order in which testimony is received.[7]

One may take the lesson here to be as follows: averaging weights of respect, or the corresponding group context which we referred to earlier, *must be sensitive to the ordering of testimonial evidence*. This is important if testimony is to be treated as *incremental evidence*, as per other kinds of evidence in Bayesian modeling. That is, past testimony on some issue should not simply be overridden by new testimony on a different issue; the sequence of reports should all contribute to an agent's final beliefs about the issues in question, such that the order in which the reports are received does not affect the final posterior distributions. We see above that fixed weights, regardless of the ordering of testimony updates, may result in different posteriors, if the partitions in question are not independent. This will not do. Therefore the group context, and thus the weights of respect assigned to experts on, say, the **B** partition, *must change*, depending on what is already known—whether testimony concerning, say, the **C** partition, has already been received.

5 Testimony as Evidence: Revisiting 'Weights of Respect'

It is all very well to outline various cases where 'group context' and thus weights of respect necessarily differ, so that linear averaging remains compatible with the Bayesian approach to belief updating. The obvious problem,

[7] Order is not important, whatever the probability updates across each partition, just in case the partitions in question are probabilistically independent. Diaconis and Zabell [5] show, moreover, that order may not be important for *some* probability updates across different partitions, even if the partitions are not probabilistically independent. They refer to this as *Jeffrey independence*—the label applies to particular partitions and particular probability updates across these partitions.

however, is that it is now very unclear what weights of respect actually amount to, and how they should be identified or measured. As noted earlier, this has been a topic of concern in the literature to date (see French [9]). The issues identified in Sections 3 and 4 only make matters worse. In both cases our recommendation was that group context must be more finely individuated than one might initially suppose; the context depends, amongst other things, on the perceived dependencies amongst group members' opinions, and now also on what has already been learnt that is not independent of the issue in question.

We do not attempt to fully answer the question of how weights of respect represent group context, and how the values of these weights can be ascertained. What follows is, rather, a suggestion that gives some minimal sense to group context, and its expression in terms of weights of respect or their functional equivalents. The basic idea is that group context may be better grasped if it is associated with a constant *evidential impact* for a given testimony report. We elaborate on evidential impact below; for now, note just that if learning experiences (here a testimony report coupled with group context) are identified with their evidential impact, then they commute. Of course, it may be that the group context and thus learning experience associated with Alice's testimony on **B** rightly differs, depending on whether or not Bob's testimony on **C** has already been received. That is all very well. The point here is just that we should understand group context in such a way that *when group context is constant, the same testimonial evidence has the same impact, and can be identified with a learning experience that commutes with other learning experiences.*

The notion of evidential impact and its relation to commutativity in the Jeffrey setting can be elaborated with reference to the work of Wagner [17], which builds on Field [8], Diaconis and Zabell [5] and Jeffrey [11]. Wagner proves the following are *sufficient* conditions for changes in probabilities across two partitions to be commutative:

Consider the following two series of probability functions due to Jeffrey-updates across the partitions **B** and **C**:

$$P \rightarrow_{\mathbf{B}} Q \rightarrow_{\mathbf{C}} R$$
$$P \rightarrow_{\mathbf{C}} S \rightarrow_{\mathbf{B}} T$$

The posterior probability functions R and T are identical if

$$\beta_P^Q(B_i, B_j) = \beta_S^T(B_i, B_j) \qquad \forall B_i, B_j$$

and

$$\beta_Q^R(C_k, C_l) = \beta_P^S(C_k, C_l) \qquad \forall C_k, C_l$$

where $\beta_P^Q(A, B) = \frac{Q(A)}{Q(B)} / \frac{P(A)}{P(B)}$ (P being the prior, Q the posterior, and β the 'Bayes factor').

Wagner's conditions for commutativity inform the notion of identical learning experiences—two cases of learning are identical if they involve the same evidential impact, or in other words, if they are characterized by the same set of Bayes factors, as defined above. Understood in this way, learning experiences commute.

Note that a change on the **B** partition to $[b_1, b_2]$ followed by a change on the **C** partition to $[c_1, c_2]$ does *not* generally amount to the same sequence of learning (only in reverse order) as a change on **C** to $[c_1, c_2]$ followed by a change on **B** to $[b_1, b_2]$. This is because the sets of relevant probability changes on the partitions, as described by the Bayes factors, differ, depending on the ordering. In the same way, we see that an averaging update on testimony with some specified set of weights may amount to different learning events, depending on the agent's priors for the partition in question.[8] Partly for this reason, the weights in the linear averaging method are very difficult to interpret.

6 An Advance on the Straight Linear Average?

The analysis of the preceding section suggests a desideratum for updating methods that would ensure that weights of respect or their equivalents are minimally comprehensible: they should encode a particular learning event, given some testimony. That is, we should understand and represent group context in such a way that the agent's shift from prior to posterior beliefs over the relevant partition yield the same Bayes factors for the same testimonial input *and group context*, $M_{\mathbf{B},G}$.

Recall the general form for testimony updating functions:

$$F_{\mathbf{B},G}: \text{ prior}, \quad M_{\mathbf{B},G} \rightarrow \text{ posterior}$$

Our new criterion is:

$$\frac{P'_0(B_i)}{P'_0(B_j)} \Big/ \frac{P_0(B_i)}{P_0(B_j)} = c_{i,j} \qquad \forall i, j$$

That is, for any two events in the **B** partition, the updating function, given a testimony profile $M_{\mathbf{B},G}$, should be such that the ratio of posteriors for the events divided by the ratio of priors is a constant.

The above criterion is satisfied by testimony updating functions that take the form:

$$P'_{0,\mathbf{B}} = \text{ normalize } [P_{0,\mathbf{B}} \times f(M_{\mathbf{B},G})]$$

where $f(M_{\mathbf{B},G})$ returns a function on **B**.

[8] For the subtleties, refer back to footnote 7.

That is, the agent's prior probability over **B** is pointwise multiplied by a function (on **B**) of the belief profile for the group of witnesses, and then normalized.[9] (The subsequent step is Jeffrey-conditionalization.) One can see that standard linear averaging (nor geometric averaging) fits this functional form.

A variant of linear averaging *does*, however, satisfy our desideratum. Consider the following:

$$\begin{aligned} P'_{0,\mathbf{B}} &= \text{ normalize } [P_{0,\mathbf{B}} \times f(M_{\mathbf{B},G})] \\ &= \text{ normalize } \left[P_{0,\mathbf{B}} \times \sum\nolimits_{i=1}^{n} w_i \times P_{i,\mathbf{B}}\right] \end{aligned}$$

where $0 \leq w_i \leq 1$ and $\sum_i w_i = 1$.

The posterior is not the linear average of all probability functions on **B**, including the agent's prior; here it equals, rather, a weighted average of the testimonial probabilities alone, *multiplied, point-wise,* by the agent's prior, and then normalised. If one wanted to defend this function, one might appeal to the popular defence of averaging given in Section 2; in this case, however, the IA and ZP properties must apply just to $f(M_{\mathbf{B},G})$ in the above expression. That is, if $f(M_{\mathbf{B},G})$ must return a probability function and must satisfy the *Independence of Irrelevant Alternatives* criterion, as well as *Zero Preservation*, then it is a weighted linear average, as specified above. Our response to Bradley in Section 3 would then also be relevant, not just for the general insights about group context, but also for the clarification of the IA criterion applied to $f(M_{\mathbf{B},G})$.

We do not, however, wish to argue vigorously for the modified linear average method outlined above.[10] There are of course many functions that have the stated form. Rather than seeking further desiderata to pinpoint a particular function for updating on testimony, it is more helpful just to highlight another property common to all functions that have this form. To this end, let us first introduce the term *defer to testimony* ; we define it here as 'changing one's beliefs to match the *aggregate* testimonial input'. The aggregate testimonial input is given by the chosen function $f(M_{\mathbf{B},G})$. The property is as follows: only

[9]Presumably this functional form is both necessary and sufficient for satisfying the desideratum, but whether it is necessary is not so obvious, and not explored here.

[10]Indeed, one might argue that the function should at least be modified to allow for the testimony having varying impact, and at the extreme, no impact at all. The following would achieve this, and indeed there would be no impact if the parameter α was set to zero (where the probability function raised to the power of α is interpreted as a point-wise operation):

$$P'_{0,\mathbf{B}} = \text{ normalize } [P_{0,\mathbf{B}} \times f(M_{\mathbf{B},G})] = \text{ normalize } \left[P_{0,\mathbf{B}} \times \left(\sum\nolimits_{i=1}^{n} w_i \times P_{i,\mathbf{B}}\right)^{\alpha}\right].$$

when the agent's prior distribution on the partition in question, say **B**, is the 'flat' distribution (i.e. equal probabilities for all events in **B**), will updating on $M_{\mathbf{B},G}$ amount to deferring to this testimonial input (in accord with the function f). One could say: only when the agent is maximally uncertain with respect to some partition, do they defer, in our sense, to testimonial input.

7 Further Issues: Sequential Updating

We have thus far been considering scenarios where an agent receives, at the one time, all the testimony on some issue **B** that they will *ever* receive. In effect, we have avoided comparing cases where testimony about some issue **B** is received en masse and cases where it is received sequentially, one individual at a time. This section briefly discusses how attending to this distinction changes the requirements on updating methods.

Presumably, if we were to address the issue of sequential testimony, we would want to apply our principle of 'same evidence and context, same impact'. Moreover, the relation between group versus sequences of individual testimony is most clear if the *impact* or the *learning event* associated with a single individual's testimony *given the context* can be separated out from the impact of others' testimony.

Such a requirement, however, amounts to an even bigger departure from the straight linear average, and the principles that justify it (IA and ZP). The aforementioned considerations impose the following form on our updating method:

$$P'_{0,\mathbf{B}} = \text{ normalize } [P_{0,\mathbf{B}} \times f_1(M_{\mathbf{B},G_{[1]}}[1]) \times \ldots \times f_n(M_{\mathbf{B},G_{[n]}}[n])]$$

where $M_{\mathbf{B},G_{[i]}}[i]$ is a single agent's probability distribution over **B**, represented by row i of the matrix, in context $G_{[i]}$ and $f_i(M_{\mathbf{B},G_{[i]}}[i])$ returns a function on **B**.

An example function with the above form is:

$$P'_{0,\mathbf{B}} = \text{ normalize } [P_{0,\mathbf{B}} \times M_{\mathbf{B},G}[1]^{w_1} \times \ldots \times M_{\mathbf{B},G}[n]^{w_n}]$$

where the parameters $w_1, \ldots, w_n$ can again be understood as weights of respect; here they need not add to one.

This example updating function is of course very similar to geometric averaging, but differences can be noted. A special case of geometric averaging *does* match our example function—the case where the principal agent is assigned weight zero and the weights $w_1 \ldots w_n$ are non-negative and summing to one. In general, however, geometric averaging, let alone linear averaging, does not match the example function, and is not consistent with the more basic functional form specified above.

8 Concluding Remarks

This paper set out to assess the normative acceptability of the weighted linear average method for updating on testimony, given that this method has some prominence in the literature. The guiding aim was to examine whether linear averaging is compatible with the Bayesian model, under a suitable interpretation of compatibility. Section 3 argued, against Bradley [2], that linear averaging *is* Bayesian-compatible, at least for single testimony updates in isolation; the key is to properly discriminate different contexts that require different 'weights of respect'. Section 4 considered multiple testimony updates on different partitions. We argued that linear averaging *can* be Bayesian-compatible, in the sense of treating testimony as incremental evidence; here again, however, the compatibility rests on finely individuating contexts that require different weights of respect.

Our investigations effectively add to an existing worry in the literature—while linear averaging may be reconciled, in a sense, with the Bayesian model, the weights of respect have no obvious meaning and so it is unclear how they can be identified or measured. This worry is articulated in Section 5. We do not offer a full solution to this problem, but propose the following: weights of respect, or more generally, the parameters of any shortcut updating method, are *minimally* comprehensible if they are identified with the testimony in question having a particular *evidential impact*, sensu Wagner.

The above criterion is satisfied by updating functions that have a particular form, described in Section 6. The straight linear average does not satisfy the functional form, but a modified linear average does, so one might borrow from the justification of the straight linear average in defending the modified rule. We leave that open. On the other hand, one might seek an even more flexible method for updating on testimony—a method that is invariant whether the testimony on some issue is received en masse or sequentially. We address this issue in Section 7, noting that it requires an even larger departure from the straight linear average.

Of course, there remains a significant issue for any proposed shortcut updating method—the *interpretation* of weights or their functional equivalents. We have suggested a basic role for these parameters, but of course there are still large questions concerning how an agent may identify different contexts that are associated with different values for these parameters. In short, the worry is that 'shortcut' updating methods are only *apparently* shortcuts, and that there is in fact much inferential work—determining dependencies amongst expert peers and the issues they report on—that is not explicitly modeled. And one could well argue that it makes more sense to model these inferences explicitly, Bayesian-style. We leave this question hanging: Given the complications identified for interpreting weights of respect or their equivalents, is there sufficient motivation for shortcut methods for updating on testimony?

Acknowledgements Many thanks to Richard Bradley for very helpful comments on an earlier draft of this paper, and to Julia Staffel and Olivier Roy, who have presented excellent comments

on this paper at the FEW and Rationality and Choice Network meetings (2011) respectively. This work was partly supported by an Internationalisation grant from the Netherlands Organisation for Scientific Research (NWO) for the 'Rationality and Decision Network'.

References

1. Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics, 4*(6), 1236–1239.
2. Bradley, R. (2007). Reaching a consensus. *Social Choice and Welfare, 29*, 609–632.
3. Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review, 116*(2), 187–217.
4. Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.
5. Diaconis, P., & Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association, 77*(380), 822–830.
6. Elga, A. (2007). Reflection and disagreement. *Noûs, 41*(3), 478–502.
7. Feldman, R. (2007). Reasonable religious disagreements. In L. Antony (Ed.), *Philosophers without God: Meditations on atheism and the secular life* (pp. 194–214). Oxford: Oxford University Press.
8. Field, H. (1978). A note on Jeffrey conditionalization. *Philosophy of Science, 45*(3), 361–367.
9. French, S. (1985). Group consensus probability distributions: A critical survey. In J. M. Bernado, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (Vol. II, pp. 183–197). Amsterdam: North-Holland.
10. Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science, 1*, 114–148.
11. Jeffrey, R. (1988). Conditioning, kinematics, and exchangeability. In B. Skyrms, & W. Harper (Eds.), *Causation, chance, and credence* (Vol. 1, pp. 221–255). Dordrecht: Kluwer.
12. Jehle, D., & Fitelson, B. (2009). What is the "equal weight view"? *Episteme, 6*(3), 280–293.
13. Kelly, T. (2005). The epistemic significance of disagreement. In J. Hawthorne, & T. Gendler Szabo (Eds.), *Oxford studies in epistemology* (Vol. 1, pp. 167–196). Oxford: Oxford University Press.
14. Lehrer, K., & Wagner, C. (1981). *Rational consensus in science and society*. Dordrecht: Reidel.
15. Shogenji, T. (2007). A conundrum in Bayesian epistemology of disagreement. Available online at www.fitelson.org/few/few_07/shogenji.pdf
16. Wagner, C. (1985). On the formal properties of weighted averaging as a method of aggregation. *Synthese, 62*, 97–108.
17. Wagner, C. (2002). Probability kinematics and commutativity. *Philosophy of Science, 69*, 266–278.