

# Bayesian Rationality in Evaluating Multiple Testimonies: Incorporating the Role of Coherence

Adam J. L. Harris and Ulrike Hahn  
Cardiff University

Routinely in day-to-day life, as well as in formal settings such as the courtroom, people must aggregate information they receive from different sources. One intuitively important but underresearched factor in this context is the degree to which the reports from different sources fit together, that is, their coherence. The authors examine a version of Bayes' theorem that not only includes factors such as prior beliefs and witness reliability, as do other models of information aggregation, but also makes transparent the effect of the coherence of multiple testimonies on the believability of the information. The results suggest that participants are sensitive to all the normatively relevant factors when assessing the believability of a set of witness testimonies.

*Keywords:* Bayesian probability, coherence, information aggregation, testimonies

Imagine sitting on a jury and three witnesses come to the stand. Each provides one piece of information: "The culprit spoke French," "The culprit wore a French football shirt," and "The culprit was waving the Tricolore flag." Given your knowledge of the distribution of French speakers, French football shirt wearers, and Tricolore flag wavers in the world, you might consider this set of information to be believable. If, however, three witnesses had said, "The culprit spoke German," "The culprit wore a French football shirt," and "The culprit was waving the Union Jack," it is likely that your belief in the veracity of this set of information would be somewhat lower. One reason why you might believe the first set of information more than the second set is that the witness reports 'hung together' better in the first information set; they were more coherent.

The role of coherence in increasing the believability of a set of information seems intuitively clear. Furthermore, there is a considerable amount of empirical research demonstrating that coherence does increase believability. For example, inconsistencies in individual prosecution witnesses' testimonies lead to lower rates of conviction as well as to lower perceptions of witness credibility (e.g., Berman & Cutler, 1996; Berman, Narby, & Cutler, 1995). Brewer, Potter, Fisher, Bond, and Luszcz (1999) further demonstrated that the consistency between reports from *different* witnesses was also perceived to be a reliable indicator of witness credibility, albeit to a lesser extent than intrawitness consistency. Moreover, Brewer et al. empirically demonstrated the validity of

using intrawitness consistency as an evaluative tool to determine a witness's credibility; objectively, consistency was a good indicator of the truth of eyewitness statements. Hence, the failure to incorporate the role of coherence into traditional models of belief revision and information aggregation represents a considerable gap in the literature.

According to the influential story model of juror decision making (Pennington & Hastie, 1993), jurors deal with trial evidence by constructing narratives. The explanatory story that most conforms to the two certainty principles, 'coverage' and 'coherence,' is the one that jurors then choose as the most likely description of the events that took place in the period addressed by the evidence. Jurors subsequently determine which verdict category best matches this story. Although Pennington and Hastie (1993) assign coherence a key role within this framework, they do not provide a way of formalizing or measuring it.

The importance of coherence to the evaluation of evidence is not limited to the courtroom; therefore a model of evidence evaluation that captures the coherence of the evidence received would be of benefit throughout cognitive psychology. In this article, we examine a recent Bayesian approach to this problem.

## The Formalization

The normative status of Bayesian probability as a formal framework for belief updating in the light of new evidence (i.e., as a prescription of how we *should* update our beliefs) is well established (e.g., Howson & Urbach, 1996), even though it has not always proved an adequate descriptive model of what people actually do (e.g., Fischhoff & Lichtenstein, 1978; Kahneman, Slovic, & Tversky, 1982; Rapoport & Wallsten, 1972; Slovic & Lichtenstein, 1971; Tversky & Kahneman, 1974; but see, e.g., Gigerenzer, Hell, & Blank, 1988; Gigerenzer & Hoffrage, 1995; Griffiths & Tenenbaum, 2006). Bovens and Hartmann (2003) demonstrated how, in Bayes' theorem, coherence, prior belief, and source reliability combine to determine how likely a set of testimonies is to be true. Two simple assumptions are required. First,

---

Adam J. L. Harris and Ulrike Hahn, Department of Psychology, Cardiff University, Cardiff, Wales.

Adam J. L. Harris was supported by an Economic and Social Research Council studentship. We thank Adam Corner, Stephan Hartmann, Andreas Jarvstad, and Jonah Schupbach for discussions, Harriet Over for comments on an earlier draft, and Louise Blackburn and Lorraine Woods for quality reproduction of experimental materials.

Correspondence concerning this article should be addressed to Adam J. L. Harris, Department of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, Wales. E-mail: harrisaj@cardiff.ac.uk

individual testimonies are assumed to be conditionally independent of one another; that is, the witnesses are conveying their own observations and have not, for example, influenced one another. Formally,  $P(R_i|F) = P(R_i|R_j, F)$ , where  $R_i$  is a report from source  $i$  and  $F$  is the fact about which they are reporting. Second, witnesses are assumed to be partially reliable; that is, they are not setting out to lie but do not necessarily report the truth. Formally,  $p > q > 0$ , where  $p$  equals the true positive rate (chance of the witness stating  $F$  is true given that  $F$  is indeed true) and  $q$  equals the false positive rate (chance of the witness stating that  $F$  is true when it is not). This also seems reasonable. If the witnesses are already known to be fully reliable (i.e., that what they say is the indubitable truth), then their reports are fully believed, and no other feature of the information set can influence the believability of that information. In addition, it is necessary for their reports to bear some relation to the truth (i.e.,  $p \neq q$ ). Otherwise, the fact that they concur can be nothing other than a coincidence.

Returning to the question of whether our Tricolore-waving, French football-shirt-wearing French speaker committed the crime, Figure 1 shows the proportions of people with such attributes in the population of possible suspects. Figure 1 also provides information relating to the co-occurrence of these attributes (their joint probability distribution). From Figure 1, one can read from the central part of the diagram that 10% of this hypothetical population are Tricolore-waving, French football-shirt-wearing French speakers, whereas only 5% speak French without wearing the football shirt or waving the flag (top left section). According to Bovens and Hartmann (2003), coherence is determined by the degree of overlap between the items of information provided. The different regions of overlap (the various  $a$  regions indicated in Figures 1 and 2) are included in the so-called probabilistic weight vector  $\langle a_0, a_1, a_2, \dots, a_n \rangle$  (abbreviated in the following as  $a_i$ );  $a_0$  captures the prior probability that all witnesses are correct,  $a_1$  the probability that all but one witness are correct (regardless of which one), and so on (see Figure 1). Bovens and Hartmann argued that the

coherence of an information set is determined by this weight vector. A maximally coherent information set is one with the weight vector  $\langle a_0, 0, \dots, 0, a_n \rangle$  where  $n$  equals the number of items in the information set, and  $a_n = 1 - a_0$ . In this set the overlap between the reports of the witnesses is perfect, and consequently either all the witnesses are correct or all the witnesses are wrong (see, e.g., Figure 2).

In addition to the weight vector,  $a_i$ , a further source of influence on the believability of an information set is the reliability of the witnesses. Bovens and Hartmann (2003) defined their reliability parameter ( $r$ ) directly from the Bayesian likelihood ratio, as  $1 - q/p$ . Given the assumptions above, Bovens and Hartmann (pp. 131–133) simplify Bayes’ theorem for the posterior degree of belief ( $P^*$ ) in the information set  $(F_1, \dots, F_n)$  having received reports  $(R_1, \dots, R_n)$ :

$$P^*(F_1, \dots, F_n) = \frac{P(R_1, \dots, R_n|F_1, \dots, F_n)P(F_1, \dots, F_n)}{P(R_1, \dots, R_n)} \tag{1}$$

to

$$P^*(F_1, \dots, F_n) = \frac{a_0}{\sum_{i=0}^n (a_i \bar{r}^i)} \tag{2}$$

where  $\bar{r}$  (“ $1 - r$ ”) equals  $q/p$ . This is a normative prescription for the updating of degree of belief in the truth of a conjunction of facts  $(F_1 \wedge F_2 \wedge \dots \wedge F_n)$  reported by multiple witnesses. The equations take into account witness reliability and prior probability judgments. Also, the influence of the coherence of the reports on the posterior degree of belief is evident from the  $a$  parameters’ interaction with reliability in the denominator of Equation 2. Different weight is thus given to information dependent on its

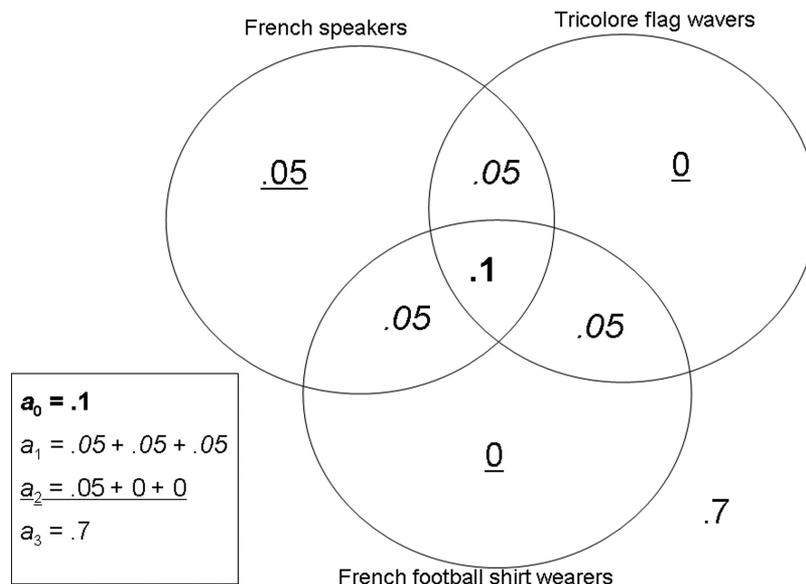


Figure 1. The co-occurrence of example attributes in the population of suspects (the joint probability distribution).

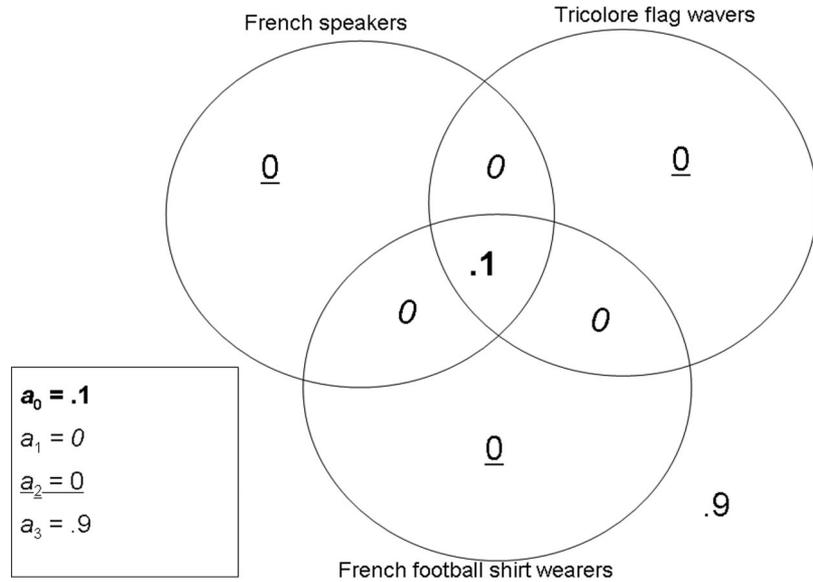


Figure 2. A maximally coherent joint probability distribution in which all (and only) French speakers wave Tricolore flags and all (and only) French speakers wear French football shirts.

degree of consistency with the other information received. Equation 2 can be illustrated with reference to the probabilistic information in Figures 1 and 2. Despite receiving reports from three witnesses with the same reliability (for this example,  $p = .8$  and  $q = .2$ , such that  $\bar{r} = .2/.8 = .25$ ), the greater coherence in Figure 2 will lead to a greater posterior degree of belief in the truth of the conjunction of the reports in that situation than for Figure 1.

If we implement Equation 2 for Figures 1 and 2, we find the following:

$$P * (F_1, \dots, F_n) = \frac{a_0}{\sum_{i=0}^n (a_i \bar{r}^i)}$$

$$P * (F_1, \dots, F_3) = \frac{.1}{(a_0 \bar{r}^0) + (a_1 \bar{r}^1) + (a_2 \bar{r}^2) + (a_3 \bar{r}^3)}$$

Figure 1:

$$P * (F_1, \dots, F_3) = \frac{.1}{(.1 \times 1) + (.15 \times .25) + (.05 \times .25^2) + (.7 \times .25^3)} = .66,$$

Figure 2:

$$P * (F_1, \dots, F_3) = \frac{.1}{(.1 \times 1) + (0 \times .25) + (0 \times .25^2) + (.9 \times .25^3)} = .88.$$

The empirical work that follows will determine to what extent participants' intuitions match the prescriptions of the Bayesian formalization.

### Comparing Participants' Intuitions With the Bayesian Predictions

We based our tests on an illustrative example given in Bovens and Hartmann (2003) in which, following a crime, witnesses provided reports of the location of a body in Tokyo. Specifically, the witnesses identify different grid locations on a visually presented map of the city. Relative to a verbal, numerical description, this visual representation of the relevant joint probability distribution provides a naturalistic format that minimizes memory and information-processing resources for participants (see Gigerenzer & Hoffrage, 1995, on problem representation). We used a range of maps, involving both two and three witnesses.

In Equation 2 above, posterior degrees of belief are calculated as a function of the  $a$  parameters and the reliability of the witnesses ( $\bar{r} = 1 - r$ ). The experimental maps are constructed such that the  $a$  parameters are directly accessible. To enable participants to make a rational judgment, and for us to make quantitative predictions, it was necessary to specify the reliability of the witnesses in each scenario. For each scenario, witness reliability was defined by first presenting participants with a map illustrating that the same witnesses had provided identical (maximally coherent) reports and telling them the police's posterior degree of belief in the truth of those reports. As participants were informed that the police knew the reliability of the two witnesses, this provided them with the information necessary to infer witness reliability.

### Method

#### Participants

Twenty-eight Cardiff University undergraduates between 18 and 38 years of age ( $Mdn = 19$ ) provided ratings of Maps A–G (displaying reports from two witnesses) in return for course credit.

A second sample of 30 students, ages 18–39 (*Mdn* = 21), provided ratings of Maps K and L (displaying reports from three witnesses) in return for £3.

*Design*

Maps A2–L2 displayed different joint probability distributions, which are shown in Table 1.<sup>1</sup> Map presentation order was quasi-randomized across all pairs for Maps A–G (a pair being each map plus its equivalent in the form of maximally coherent information) with seven orders of presentation, each of which was seen by 4 participants. These orders were constructed such that each pair of maps appeared in each serial position once and one pair did not always immediately precede another. For Maps K and L, the order of map presentation was counterbalanced.

*Materials and Procedure*

Eighteen maps were designed, each portraying differentially coherent information from a number of witnesses, in which each witness is reporting on the possible location of a body. Nine of the maps (A1–L1) were designed to contain the information in maximally coherent form (such that  $a_0$  was the same as in the corresponding test map, but  $a_1, \dots, a_{n-1} = 0$ ). The remaining nine (Maps A2–L2; see Figure 3 for examples of the maps used) provided the actual test stimuli. Reports from either two or three witnesses were illustrated on the maps with both shading and labeling. Different colors were used for areas that different numbers of witnesses agreed upon as possible locations of the body. The maps used were downloaded from the Internet (<http://www.map-of-spain.co.uk/maps-of-spain/madrid/iMadridG-med.jpg> and [http://images.worldres.com/search/lang1/map\\_edinburgh.png](http://images.worldres.com/search/lang1/map_edinburgh.png)) and were simplified maps of Madrid (Maps C–L) and Edinburgh (Maps A and B).

The experimenter handed participants a pair of maps with each page of an experimental booklet (which contained the written scenarios). Participants were always handed the pair of maps with the maximally coherent map on top. For each written scenario, participants first read:<sup>2</sup>

A man has gone missing in Madrid and is presumed murdered. The police are now searching for the body.

Table 1  
*Joint Probability Distributions Represented by the Different Maps, Reliability of the Witnesses (r) in These Maps, and Predicted Ratings*

| Map | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $r$ | Predicted rating |
|-----|-------|-------|-------|-------|-----|------------------|
| A2  | .01   | .10   | .89   | .00   | .96 | 12.65            |
| B2  | .80   | .10   | .10   | .00   | .16 | 16.76            |
| C2  | .12   | .60   | .28   | .00   | .84 | 10.92            |
| D2  | .10   | .02   | .88   | .00   | .86 | 16.66            |
| E2  | .01   | .70   | .29   | .00   | .96 | 4.99             |
| F2  | .10   | .20   | .70   | .00   | .86 | 14.11            |
| G2  | .12   | .22   | .66   | .00   | .84 | 14.12            |
| K2  | .09   | .00   | .61   | .30   | .83 | 16.38            |
| L2  | .09   | .33   | .28   | .30   | .83 | 11.45            |

They KNOW that the body must be in one of 100 grid squares that they have divided the city into. Furthermore, they have received independent tip offs from 2 sources as to the location of the body. Both sources are known by the police to be less than fully reliable.

In the first situation, both sources identified the same possible locations of the body, and these are shaded red on map C1.

After receiving this report the police indicated their belief that this information was correct using the scale below, where their response is shown:

On a scale of 0–20, the police had circled 17 in answer to the question: “How likely do you think it is that the body lies in the area of the map shaded red?”<sup>3</sup>

The second half of the page was introduced with the words “Now consider the following scenario”: The first two paragraphs of the second scenario were identical to those for the first scenario, but the third paragraph was replaced by the following two:

This time, each source (1 and 2) reports that the body is in a different area of the city, but there is some overlap between their reports.

On map C2, the areas that source 1 has identified as possible locations of the body are shaded green, whilst those identified by source 2 are shaded yellow. Crucially, the areas reported by the 2 sources overlap, and this area is shaded red.

Participants were then asked, “How likely do you think the police should think it is that the body lies in the area of the map shaded red?” Participants made their responses on a scale from 0 to 20.

Results

Prior to data analysis, the data from 3 participants in Sample 1 and 1 in Sample 2 were excluded for their failure to follow simple task instructions.

The inclusion of the police’s posterior degree of belief rating for the maximally coherent maps enabled precise quantitative predictions to be made of the believability ratings for each map. As noted above (Equation 2), posterior degree of belief corresponds to

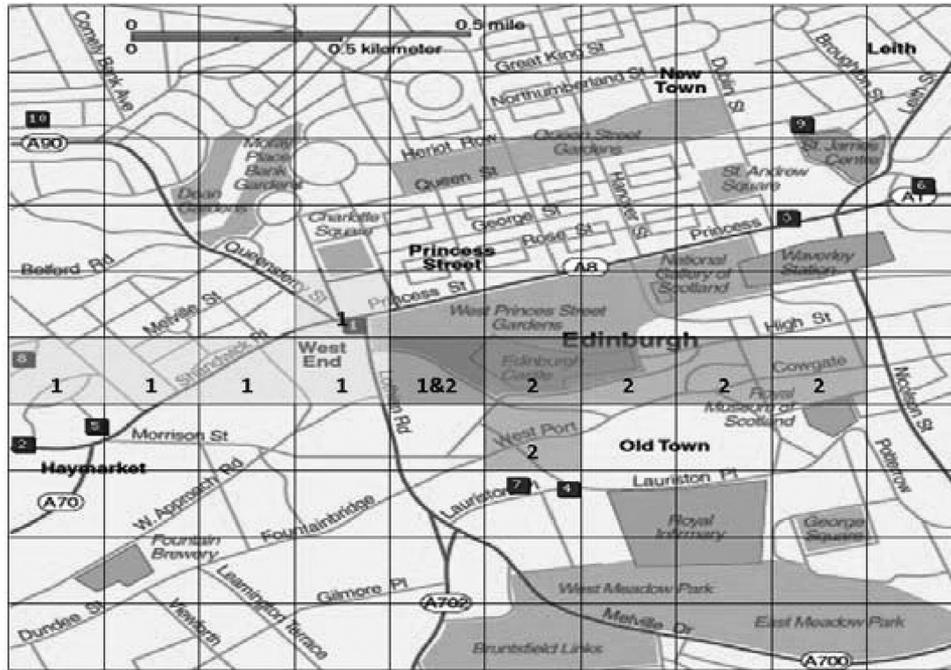
$$P * (F_1, \dots, F_n) = \frac{a_0}{\sum_{i=0}^n (a_i \bar{r}^i)}$$

where  $n$  is the number of reports received. Knowing the value of  $P^*$  for the maximally coherent set of information, we used

<sup>1</sup> The probability distributions represented in Maps A, B, K, and L also enable comparisons between competing measures of coherence suggested in the philosophical literature (Bovens & Hartmann, 2003; Fitelson, 2003; Olsson, 2002). This issue is not pursued further here.

<sup>2</sup> The text was identical for the Edinburgh maps, but different (more distinctive) colors were used, and the word *Madrid* changed to *Edinburgh*. Maps K and L displayed the reports of three witnesses; the wording was changed appropriately.

<sup>3</sup> For Maps K and L, 19 had been circled by the police. This was to maximize the difference in predicted responses between these maps.



A2



L2

Figure 3. Two black-and-white examples (A2 and L2) of the maps used. Different colored shading and appropriate numbers inside the shaded grid squares illustrated how many and which witnesses had indicated a particular grid location as a possible location of the body.

this equation to derive the reliability parameter ( $r$ ) for the witnesses in each map (see Table 1). With these values, we were able to calculate the Bayesian predictions for the test maps using the same equation. The probabilistic predictions were transformed into predicted responses on a scale of 0–20 (through multiplying by 20) and are shown in the right-hand column of Table 1.

Having thus calculated the predicted ratings, we can provide a strict test of the rationality of participants' ratings with respect to the Bayesian standard. The correlation between the predicted and observed data (averaged across participants for each map) was significant,  $r(7) = .92, p < .001, r_{\text{adj}} = .91$  (Howell, 1997, p. 240), indicating that 83% of the variance in participants' ratings of the different maps was explained by the Bayesian model (see Figure 4). This is the main statistic for evaluating the model because it takes all maps into account.

If we restrict our analysis to Maps A–G, however, it is possible to conduct model comparisons to both group and individual data.<sup>4</sup> In the corresponding group-level analysis for these maps, the Bayesian model accounts for 91% of the variance in participants' mean ratings across the maps. In the individual-level analysis, we calculated a correlation coefficient for each participant and then averaged these coefficients across participants. This yielded a mean individual correlation value of .52. As Wallsten, Budescu, Erev, and Diederich (1997) show, where participants' judgments are subject to some degree of random error or noise, a group-level average will be closer to the true underlying values. Hence, the observed difference between the two types of averaging method itself provides some support for the hypothesis that participants' judgments are (somewhat) noisy estimates of the Bayesian predictions.

In line with the above conclusion, we found further that when examining individual responses across all maps (A–L), a quarter of all responses (55 out of 229) made were within one response unit of the predicted value, such that were the critical value 12.8, these responses were either 12 or 13, whereas half of all responses (114) were within two response units of the predicted value.

Notably, there were no free parameters that could be 'tweaked' to match the experimental data. The close correspondence between the predicted and observed data demonstrates participants' sensitivity to the probabilistic parameters already inherent in the scenario.

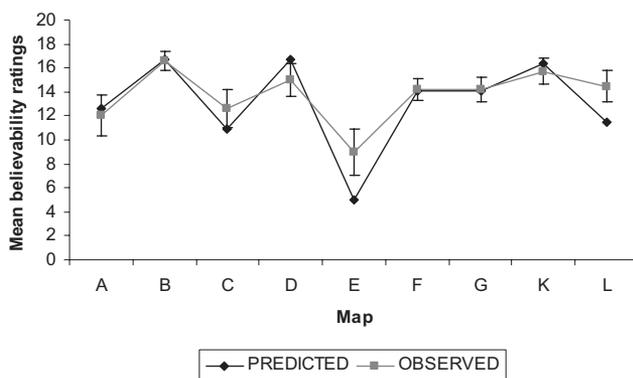


Figure 4. Predicted believability ratings and observed believability ratings across all nine maps. Error bars are  $\pm 95\%$  confidence intervals.

## Discussion

Our participants were able to process and aggregate complex information in a manner consistent with the prescriptions of Bayesian probability. The high correlations between participants' believability ratings and the Bayesian predictions were especially impressive given that participants had to combine prior probabilities, the coherence of the reports, and witness reliability.

Traditional averaging models of information aggregation (e.g., Anderson, 1971; Birnbaum & Stegner, 1979; Hogarth & Einhorn, 1992) have never been brought to bear on issues such as coherence because they require numerical quantities—a scale value of some sort—for their application. To illustrate, the average of “waves the Tricolore,” “wears a French football shirt,” and “speaks French” as qualitative statements is simply undefined. As seen here, probabilities can provide a common currency with which not only quantitative but also qualitative units can be combined. Consequently, once a probabilistic formalization of a problem has been derived, some form of averaging model can then also be defined, though it will not necessarily be one that has been devised for and used in other tasks. Traditional averaging models bear no straightforward relationship to our task, so they cannot readily be applied here. Thus there are no candidate models in the extant literature. Given the popularity of such models (see, e.g., Clemen, 1989, and references therein), however, we devised a plausible averaging model for our maps for comparison with the Bayesian account.

A grid square that all witnesses agreed was a possible location of the body was given a ‘support’ value of 1. A grid square that not all witnesses agreed on as a possible location was given a support value that corresponded to the number of witnesses stating that location divided by the total number of witnesses. To determine the likelihood of the body being in a certain location, we calculated and normalized the total support for that location with respect to the total support for all areas of the map. Overall, this averaging model accounted for 50% ( $r_{\text{adj}}^2$ ) of the variance in the mean responses to the nine maps (Figure 5). The reasonable performance of this model is to be expected, as it is systematically related to the Bayesian model, which also takes into account the degree of overlap between different reports. It does not, however, account for as much variance as the Bayesian model and seems to have particular difficulty in accounting for the relationship between the three witness maps (K and L) and the two witness maps (A–G). In line with Glover and Dixon's (2004) metric, the experimental data were 111 times more likely to occur under the Bayesian model than under the averaging model.<sup>5</sup> This suggests that participants were sensitive to all the normatively relevant factors (prior belief, reliability, and coherence) when making their judgments. In the corresponding group and individual average correlations on Maps A–G, the averaging model captured 84% and 47% of the variance,

<sup>4</sup> Maps K and L were evaluated by a different set of participants, so they cannot be included in this analysis.

<sup>5</sup> Spearman rank order correlations for the fits between the models and the data were  $r(7) = .83, p < .05$ , for the Bayesian model and  $r(7) = .78, p < .05$ , for the averaging model. By this analysis, the Bayesian model is 2.8 times more likely to be the true underlying model of the data than the averaging model is (Glover & Dixon, 2004). Given, however, that the Bayesian model makes precise quantitative predictions, the parametric correlations reported in the main text are the most appropriate analyses.

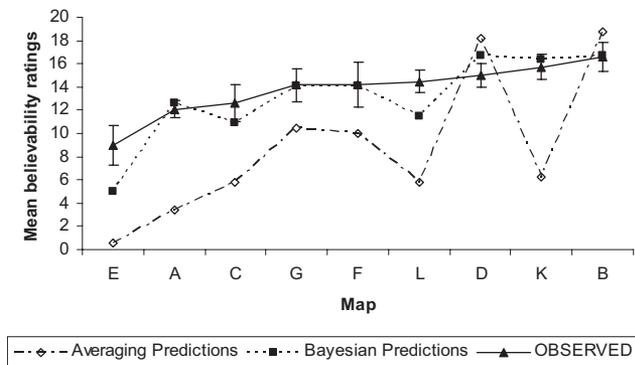


Figure 5. A comparison of the fits of a simple averaging model and the Bayesian model with the observed ratings. Maps are arranged in order of increasing observed ratings. Error bars are  $\pm 95\%$  confidence intervals.

respectively (an improvement that emphasizes the model's difficulty with Maps K and L).

The better predictive performance of the Bayesian model over the averaging model is even more impressive when one considers the nature of the reliability manipulation. Participants were *not* told that they were being shown a set of reports provided by very reliable witnesses in Map A2 and very unreliable witnesses in Map B2, for example. Participants were simply told that the witnesses were less than fully reliable, and this information was exactly the same across all maps. On the Bayesian account, differences in reliability are inferred from the anchoring map (the maximally coherent case) because there is an equation that systematically relates reliability, overlap, and prior probability. Without this equation, there is no way of implicitly inferring differential reliabilities. Consequently, there is no obvious way to incorporate reliability into the averaging model.

In summary, averaging models do not seem a promising approach for this domain. As noted, they require a translation of qualitative information into numbers to be applicable. Probabilities can provide such a translation; however, the impressive data fit for the Bayesian model and the systematic relationship between the two models suggest that people are already sensitive to all the quantities that figure in a Bayesian approach. This makes it hard to see how averaging could provide a more suitable framework for this problem.

In an alternative to the approach pursued here, Thagard (1989, 2000, 2005) has proposed that the juror's fact-finding process is best modeled (descriptively and normatively) in terms of *explanatory coherence*, which is explicitly proposed as an alternative to Bayesian probability (e.g., Thagard, 2005, p. 312). Despite Thagard's claim that explanatory coherence is normatively (as well as descriptively) the more appropriate theory, his argument is made entirely on descriptive grounds (a point also noted by Papineau, 1989), primarily highlighting the difficulties of interpreting and implementing the necessary conditional probabilities for a Bayesian analysis. The data reported in this article directly address this critique by demonstrating that people can be rational with respect to a Bayesian probabilistic norm. Furthermore, the benefits of a probabilistic normative standard for reasoning within the courtroom can be illustrated with reference to the issue of reasonable doubt. Blackwell's maxim (see, e.g., Nagel, Lamm, & Neef, 1981)

states that it is better for 10 guilty men to go free than for one innocent man to be convicted; in other words, the probability of guilt should be greater than 90% before a guilty verdict is passed. More generally, Bayesian decision theory provides a simple formal framework for deriving burdens of proof such as the one embodied in the standard of reasonable doubt from the potential consequences associated with a decision (see also Hahn & Oaksford, 2007).<sup>6</sup>

Pennington and Hastie (1993), too, have been skeptical about the psychological utility of Bayesian belief updating, but the results reported in this article enhance the appeal of a Bayesian formalization of the juror's fact-finding task. Especially given the pragmatic 'sleuthlike' nature of the materials used, the good model fits provide support for the contention that the Bayesian framework could provide a suitable normative standard for the fact-finding process, which jurors could realistically aspire to. In addition, the features of an information set that are captured within the Bayesian account (namely witness reliability, prior belief, coherence of testimonies) seem, intuitively, to match those that we would wish to capture in an analysis of the believability of testimony. The present study may, therefore, form the foundations for a research program integrating powerful Bayesian computational models with well-supported process models of juror decision making, such as Pennington and Hastie's story model, to determine conditions for maximizing the efficiency of jurors' fact-finding.

The empirical findings presented here represent the first step in this line of research, which appears to have considerable potential for development. For example, in the scenarios presented in our experiments, regardless of the proportion of overlap between the witnesses' reports (which is captured by a consideration of the degree of coherence), it was always possible for all witnesses to be correct. In the real world, however, we encounter evidence that is not only more or less coherent (as in our experiment) but also evidence that is simply *incoherent*, that is, claims that cannot simultaneously be true (e.g., the defendant was in Brazil when the crime occurred; the defendant was in Japan when the crime occurred). Lagnado and Harvey (2008) have proposed that when engaging in complex reasoning tasks (i.e., those in which people receive multiple reports, of which the truth of some might preclude the truth of others, such as the juror's fact-finding task), people group propositions into coherent sets (i.e., those that are not mutually exclusive). Within the present formalization, the believability of the two inconsistent information sets could then be compared to determine which of the two is more likely and indeed how much more likely it is.

The particular version of Bayes' theorem examined in this article required two explicit assumptions that seem reasonable (the conditional independence of the testimonies and imperfect source reliability) and are supported in the experimental materials. The conditional independence assumption captures those situations in which multiple witness testimonies are most informative. However, in situations in which testimonies are not conditionally independent, the information recipient must be sensitive to the

<sup>6</sup> Although Thagard (2004) cites probabilistic blindness to the emotional character of doubt as a limitation for a probabilistic approach, he does not address how an appropriate standard of reasonable doubt is devised from his theory.

conditional dependencies present in order to maintain rationality. Within the Bayesian framework, these relationships between evidence items are best modeled with hierarchical networks (e.g., Bovens & Hartmann, 2003; see also Pearl, 2000; Schum, 1994). Bovens and Hartmann (2003) use Bayesian networks to model situations in which multiple testimonies are received from the same witness (and are therefore not conditionally independent). Future work in this area could provide a means for assessing the degree to which inconsistencies in a single witness's testimony (e.g., Berman & Cutler, 1996; Berman et al., 1995; Brewer et al., 1999) should influence perceptions of that witness's credibility. Within the Bayesian framework, recognition of the interdependencies between evidence items allows rational belief updating not only with respect to the truth of the hypothesis but also with respect to the reliability of the witnesses. In keeping with this, Lagnado (in press) has provided initial evidence that people are sensitive to intricate conditional dependencies between evidence items in legal scenarios.

In conclusion, the excellent, parameter-free fits obtained between behavior and model in our study demonstrate that people can adhere to the prescriptions of a Bayesian theory of testimony believability. Future research should take up the challenge of harnessing the rationality observed in this experiment for those practical situations in which the accuracy of judgment is of most importance. Finally, the results presented here provide the first empirical evidence to suggest that the important goal of providing a quantitative measure of the coherence of witness testimony may be achievable. Furthermore, such a measure seems likely to exist within the established normative framework of Bayesian probability.

## References

- Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review*, 78, 171–206.
- Berman, G. L., & Cutler, B. L. (1996). Effects of inconsistencies in eyewitness testimony and mock-juror decision making. *Journal of Applied Psychology*, 81, 170–177.
- Berman, G. L., Narby, D. J., & Cutler, B. L. (1995). Effects of inconsistent eyewitness statements on mock-juror's evaluations of the eyewitness, perceptions of defendant culpability and verdicts. *Law and Human Behavior*, 19, 79–88.
- Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, 37, 48–74.
- Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford, England: Oxford University Press.
- Brewer, N., Potter, R., Fisher, R. P., Bond, N., & Luszcz, M. A. (1999). Beliefs and data on the relationship between consistency and accuracy of eyewitness testimony. *Applied Cognitive Psychology*, 13, 297–313.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5, 559–583.
- Fischhoff, B., & Lichtenstein, S. (1978). Don't attribute this to Reverend Bayes. *Psychological Bulletin*, 85, 239–243.
- Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, 63, 194–199.
- Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 513–525.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684–704.
- Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11, 791–806.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773.
- Hahn, U., & Oaksford, M. (2007). The burden of proof and its role in argumentation. *Argumentation*, 21, 39–61.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1–55.
- Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.
- Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: Open Court.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Lagnado, D. A. (in press). Thinking about evidence. In P. Dawid, W. Twining, & M. Vasilaki (Eds.), *Evidence, inference and enquiry*. Oxford, England: Oxford University Press.
- Lagnado, D. A., & Harvey, N. (2008). The impact of discredited evidence. *Psychonomic Bulletin & Review*, 15, 1166–1173.
- Nagel, S., Lamm, D., & Neef, M. (1981). Decision theory and juror decision-making. In B. D. Sales (Ed.), *The trial process* (pp. 353–386). New York: Plenum Press.
- Olsson, E. J. (2002). What is the problem of coherence and truth? *Journal of Philosophy*, 94, 246–272.
- Papineau, D. (1989). Probability and normativity. *Behavioral and Brain Sciences*, 12, 484–485.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 192–221). Cambridge, England: Cambridge University Press.
- Rapoport, A., & Wallsten, T. S. (1972). Individual decision behavior. *Annual Review of Psychology*, 23, 131–176.
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.
- Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of human information processing in judgment. *Organizational Behavior and Human Performance*, 6, 649–744.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435–502.
- Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.
- Thagard, P. (2004). What is doubt and when is it reasonable? In M. Ezcurdia, R. J. Stainton, & C. Viger (Eds.), *New essays in the philosophy of language and mind* (pp. 391–406). Calgary, Alberta, Canada: University of Calgary Press.
- Thagard, P. (2005). Testimony, credibility, and explanatory coherence. *Erkenntnis*, 63, 295–316.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, 10, 243–268.

Received February 6, 2009

Revision received April 28, 2009

Accepted May 7, 2009 ■