A Cue Imputation Bayesian Model of Information Aggregation

Jennifer S. Trueblood, George Kachergis, and John K. Kruschke
{jstruebl, gkacherg, kruschke}@indiana.edu
Cognitive Science Program, 819 Eigenmann, 1910 E. 10th St.
Bloomington, IN 47406 USA

Abstract

Decision makers are sometimes faced with aggregating advice from multiple advisers without knowing what information is driving each adviser's opinion. Following Budescu and Yu (2006, 2007), we conducted an experiment in which participants first learned to estimate the probability of a disease based on multiple test results. Next, subjects made the same judgments solely on the basis of probabilities given by multiple advisers who may have only received partial information. Experimental results confirm previous findings that decision makers give extreme estimates when advisers are in agreement and compromise estimates when advisors are in disagreement. Unlike previously proposed models that can only account for extreme or compromise estimates but not both, we develop a new Bayesian model that explains both types of judgments. This model provides a rational explanation of information aggregation by assuming that decision makers use the probability estimates of advisers to infer underlying data before making probability judgments.

Keywords: information aggregation, decision making, Bayesian model

Introduction

In many real world judgment problems, decision makers often make inferences by aggregating information from multiple outside sources. For example, a juror might be asked to determine the probability that a defendant is guilty based on the testimony of expert witnesses such as a medical examiner and a forensic scientist. In this paper, we discuss inference problems in which a decision maker must aggregate judgments from multiple expert advisers whose judgments might be based on partial information. Our goal is to develop a probabilistic framework that can account for the aggregated judgments of decision makers. Specifically, we propose a new Bayesian model to explain aggregated judgments and compare our model to a previously proposed Bayesian model (Budescu & Yu, 2006, 2007) and a weighted averaging model. The newly proposed Bayesian model is able to account for both extreme and conservative judgments unlike previous models of information aggregation.

The modeling and experimental work described in this paper was motivated by a two stage study of information aggregation conducted by Budescu and Yu (2006). In this experiment, subjects first completed a learning stage in which they learned to distinguish between two diseases (A and B) based on a set of six cues (i.e., medical tests) with binary values (i.e, positive or negative test results). Subjects were told that the cues were equally valid, not perfectly diagnostic, and could be correlated. In the second stage of the experiment, some subjects were selected to be advisers and other subjects were selected to be decision makers. The advisers provided the decision makers with probability estimates of disease A based on a set of cues. In some situations, the advisers saw only a portion of the six cues. In these cases, not all advisers saw the same subset of cues. Then, the decision maker was asked to use the advisers' judgments to provide a probability estimate for disease A. The decision maker was told which binary cues each adviser saw; however, the decision maker was not privy to the cue values (i.e., positive or negative). The number of advisers giving advice was either two or three.

The experimental results indicated that in some situations decision makers gave estimates more extreme than both advisers' values. In general, when the advisers displayed a high level of agreement, decision makers produced extreme answers. On the other hand, when the advisers' displayed disagreement, the decision makers seemed to average the advisers' estimates. A simple Bayesian model presented by Budescu and Yu predicts that decision makers will always produce extreme estimates. Since this was not found to be the case, Budescu and Yu (2006) claimed that decision makers are not always Bayesian. However, we hope to show that a novel Bayesian model can account for compromise judgments when there is a discrepancy among advisers' probability estimates and extreme judgments when advisers display a high level of agreement.

Experiment

We conducted an experiment with a learning stage followed by an information aggregation stage, similar to Experiment 2 in Budescu and Yu (2007). Participants first learned to diagnose, using the binary results of three tests, the probability that a fictitious disease was present in a series of patients. The probability of a disease given a set of test results was determined by a causal graph. Specifically, it was assumed that a disease caused hidden intermediate states that determined the test results. For simplicity, we assumed that the intermediate states were binary like the test results. For example, lung cancer causes malignant tumors resulting in a positive lung biopsy. Since Budescu and Yu allowed for tests to either be correlated or not, we used the causal graph structure to represent correlation among tests. We denoted diagnostic tests that were driven by common underlying causes as correlated. For example, a malignant tumor causes both a positive lung biopsy and a positive PET scan. Diagnostic tests that were driven by different causes were called uncorrelated. In our lung cancer example, a malignant tumor causes a positive lung biopsy and a high white blood cell (WBC) count produces a positive blood test.

After the training stage, participants were given a series of trials on which two experts gave the participant the probability a patient has the disease. Participants could see which test results each expert was privy to, but could not see the results, themselves. The probability given by each expert was produced from the causal graphs, and thus depends on the number and values of test results they were privy to. Faced with advice from two experts, participants may use a number of strategies to aggregate this information. Given that the advisers have overlapping information in cases when they saw some of the same test results, and unique information in cases when advisers saw disjunct (and perhaps conflicting) test results, a rational strategy would be for participants to infer the union of the results that the advisers saw, and thence produce a probability estimate as they did in the learning phase.

Subjects

71 undergraduates at Indiana University participated in the experiment for course credit.

Stimuli & Procedure

Participants were instructed that they would be learning to diagnose patients with a fictional disease (e.g., 'nomitis') on the basis of results from three tests. As in Budescu and Yu (2007), participants were told that each test is equally diagnostic, that the probability of the disease in the patient population is 0.5, and that results of the medical test may be correlated. A high prevalence of disease was used in order to match the experimental paradigm of Budescu and Yu (2007). Subjects were told that there was an epidemic in order to make the high base rate of the disease seem plausible.

On each of 88 training trials, participants were shown the binary results ('+' or '−') of three tests ('A', 'B', or 'C'), and asked to indicate the probability (0-100) that a patient with these test results has the disease. Probabilistic feedback was provided on each trial: participants were told that this particular patient did or did not have the disease (proportional to the predetermined disease probabilities), and were also told whether or not a patient exhibiting these results would typically have the disease.

Appearing after the initial 88 trials, an additional 19 training trials showed only 0, 1, or 2 test results, and no feedback was given after participants responded. These trials introduced the possibility that only a subset of the test results may be seen, as is true in many of the information aggregation trials.

For 36 of the participants, test results were uncorrelated, whereas test results for the remaining 35 participants were correlated.

In the second stage of the experiment, participants were told that they had been promoted to a supervisory position in which they would continue to judge the probability a patient has the same disease. However, instead of looking at test results themselves, participants would see the probabilities given by two advisers who have had the same training they have had, but who may see only some of the test results (e.g., Figure 1, left). Participants were instructed to keep in mind that each adviser's advice would be based on both the number of results they were privy to, and the outcomes of the results. As a reminder, for the first 12 adviser trials, decision makers also saw the test results the advisers saw (e.g., Figure 1, right). These trials and the remaining 44 adviser trials were identical for decision makers in both conditions, except that the particular probabilities given as advice depended on the condition (based on the two types of causal graphs: correlated and uncorrelated). The 44 trials included all combinations of strength (i.e., number of agreeing test results) and amount (i.e., number of test results seen: 0–3) of evidence seen by an adviser, as well as various combinations of the amount seen by both advisers.

Trials could vary by the amount of overlapping (i.e., redundant) evidence seen by the two advisers: all of the results seen by the advisers may overlap, none of them may overlap (e.g., Figure 1, right), or they may partially overlap (e.g., Figure 1, left). On a given trial, when advisers are privy to non-overlapping test results (e.g., A and B in Figure 1, left), and the two advisers agree on the diagnosis (both less than 50, or both greater), the decision maker, inferring what each test result is, should normatively *extremify*. In other words, the decision maker should respond further from 50 than either adviser. The reason is that, by inferring the test results, the decision maker has more information than either adviser, and the decision maker knows that this inferred information is stronger than the individual pieces of information seen by each advisers. On other trials, when one adviser sees test results that are a subset of the other adviser's, the decision maker should *match* the rating of the adviser who saw the most, because the adviser who saw the subset of tests provides no additional information. Finally, in cases where advisers saw non-overlapping results but disagree on the diagnosis (e.g., Figure 1, right), decision makers should *compromise*, and give a rating somewhere between what the advisers gave. Like the case of extremifying, the decision maker is able to infer more information than either adviser. However, in this scenario, the inferred information is contradictory and leads to a compromising estimate.

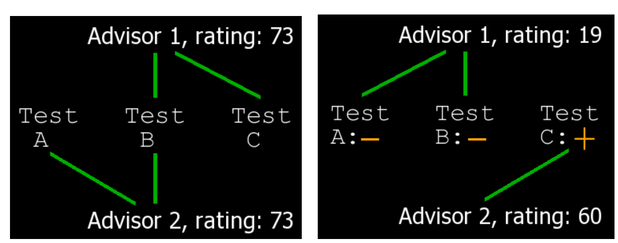

Figure 1: Information given on adviser trials: partial overlap on which decision makers should extremify (left), and no overlap on which decision makers should compromise (right). Actual test results (+/-) were only given on the first 12 adviser trials.

Results & Discussion

The best theoretical accuracy during training is 0.68 for both conditions, possible only if a participant gives a probability rating in accord with the more likely outcome for the test results on all 88 complete trials. In the last half of training, accuracy in the uncorrelated tests condition (M = .660) is higher than accuracy in the correlated tests condition (M = .615). In a Bayesian estimate of the difference between groups, the mean of the posterior was 0.045, with the 95% highest density interval extending from 0.0283 to 0.0613, and with 100% of the posterior distribution falling above 0.02 (Kruschke, 2011). Not surprisingly, correlated test results confuse learners somewhat. Since we are primarily interested in the way that experts aggregate advice in the second stage, we chose to exclude participants who did not become experts (i.e., those who were more than a standard deviation below the group mean of their condition). In the uncorrelated condition, this criterion ($M - \sigma = .57$) excludes 6 participants, and in the correlated condition ($M - \sigma = .52$) excludes 7 participants.

An additional participant was excluded from the correlated condition because they responded '50' on every adviser trial. Thus, the adviser trials of 30 participants in the uncorrelated condition and 27 participants in the correlated condition were analyzed. To evaluate aggregation performance, we examined the proportion of times each participant extremified, compromised, and matched relative to the number of times they should have shown each of these behaviors according to our normative model of the task. We found that the graph structure had no notable effect on these proportions, so Table 1 shows the aggregate performance of all 57 participants. For each type of normative behavior, indicated by rows in Table 1, the most often observed behavior was the normative one. For example, when the normative behavior was to extremify, 44% of responses were extremifications, which is more than any other type of response.

Table 1: Adviser trial behavior by normatively correct behavior. Each cell shows p(observed | normative).

	Observed			
Normative	Extremify	Compromise	Match	Other
Extremify	.44	.12	.14	.29
Compromise	.19	.68	.13	0
Match	.32	.21	.40	.07
Overall	.32	.34	.22	.12

Models

Using the data collected from the new experiment, we compared three different aggregation models, a weighted mean log-odds model (WMLO), a naively adjusted Bayesian model, and the new cue imputation Bayesian model. Before we discuss the model comparison, we describe each of the three models in detail. The WMLO and naively adjusted Bayesian models were two of the many different models examined by Budescu and Yu (2006). We selected these models to use in our comparison because the WMLO and adjusted naive Bayesian model fit the data from Budescu and Yu well from a global point of view. Thus, we consider these models as good competitors for testing our new Bayesian model.

An Averaging Model

The weighted mean log-odds model is an averaging model which accentuates differences between extreme probabilities. By using log-odds, the high and low probabilities are stretched out before they are averaged. Letting E be the target event (i.e., presence of a disease), we say that adviser j provides a probabilistic forecast $p_j = p(E|C_j)$ based on the set of cues C_j. The WMLO model is given by the following formula

$$WMLO = \frac{e^{\beta}}{1 - e^{\beta}} \tag{1}$$

where

$$\beta = \frac{1}{\sum_{j=1}^{J} n_j} \cdot \sum_{j=1}^{J} (n_j \cdot ln \frac{p_j}{1 - p_j}) \tag{2}$$

and where J is the number of advisers and n_j is number of cues seen by adviser j. This model can out perform a simple averaging model because it can better account for extreme judgments; however, it cannot extremify. In other words, the model cannot respond further from 50 than either adviser. Thus, if we compared the model predictions to the normatively correct behavior as was done in Table 1 for the behavioral data, the cell of the table indicating an extremifying prediction when extremifying is the normatively correct behavior would be 0. Also note that this is a parameter free model.

Naively Adjusted Bayesian Model

We begin our discussion of the two Bayesian models of information aggregation by describing the naively adjusted Bayesian model (Budescu & Yu, 2006). This model is an extension of a naive Bayesian model that calculates the posterior probability of an event, E, assuming conditional independence of the advisers' forecasts. By the conditional independence assumption and Bayes' rule, the posterior probability is given by

$$p(E \mid \bigcap C_j) = \frac{\prod_{j=1}^{J} p_j / p(E)^{J-1}}{\prod_{j=1}^{J} p_j / p(E)^{J-1} + \prod_{j=1}^{J} (1 - p_j) / p(\overline{E})^{J-1}} \tag{3}$$

where J denotes the number of advisers and p_j is defined as above. As noted by Budescu and Yu, this is very similar to the aggregation model by Bordley (1982) and the aggregation rule by Genest and Schervish (1985).

This naive Bayesian model assumes that decision makers treat the judgments from each adviser as perfectly reliable. In order to allow decision makers to adjust their judgments based on different assumptions of reliability, Budescu and Yu incorporated a discounting function into the model. Specifically, before the probability judgments are aggregated, they are discounted according to the model by Karmarkar (1978) given by

$$p'_j = \frac{p_j^\lambda}{p_j^\lambda + (1-p_j)^\lambda} \quad (4)$$

where $0 < \lambda < 1$ is a parameter associated with the level of discounting. When $\lambda = 1$, no adjustment is made and the probability judgment is considered perfectly reliable. However, when $\lambda = 0$, the probability is transformed into 0.5, and the probability judgment is considered not diagnostic. The overall effect of this model is shrinkage of extreme probabilities before aggregation. The naively adjusted Bayesian model can provide a better account of conservative probability estimates than the simpler naive Bayesian model; however, it cannot compromise. In other words, the model cannot give a rating somewhere between the advisers' ratings. If we compared the model predictions to the normatively correct behavior as was done in Table 1 for the behavioral data, the cell of the table indicating a compromise prediction when compromising is the normatively correct behavior would be 0.

A Cue Imputation Bayesian Model

Because both the information aggregation study by Budescu and Yu and our new experiment consisted of a learning stage and an evaluation stage, we include in our Bayesian description of the task a model of the learning process. None of the models considered by Budescu and Yu included a description of the learning stage of the task.

Begin by letting X_t be the data seen on trial t. For our purposes, X_t is a set of dichotomized test results related to diagnosing the potential disease. For the case of three tests as in the experiment described above, we might have $X_t = \{+,+,-\}$. Let the hypothesis space, H, be the set of causal graphs (or Bayesian networks) representing the relationship between the disease and the test results. It is assumed that a disease causes hidden intermediate states (labeled S below) that drive the test results (labeled T below).

For simplicity, consider the case when there are only two diagnostic tests. In this case, the correlated causal graph shows that a disease causes a single intermediate state which drives the results of the two tests (see Figure 2 left). On the other hand, the uncorrelated causal graph shows that a disease causes two intermediate states which each drive the results of corresponding tests (see Figure 2 right). We can represent causal graphs by conditional probabilities. For example, we might purpose Table 2 for our uncorrelated causal graph.

The Learning Stage In the learning stage, the model learns the probability that the tests are correlated. Let $h \in H$ represent a causal graph. In the case with two tests, we might assume that there are only two causal graphs as shown in Figure 2. Of course, a larger hypothesis space and more complicated graphical structures can be used. By Bayes' Rule, we define the posterior probability of each hypothesis (i.e. causal structure) on trial t by

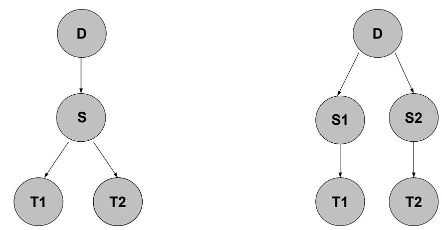

Figure 2: Causal graphical models for two correlated tests (left) and for two uncorrelated tests (right).

Table 2: Example of Conditional Probabilities for Uncorrelated Tests

Diseases	Intermediate States	Tests
p(D= +) = .5	p(S_i= + | D= +) = .8	p(T_i= + | S_i= +) = .9
p(D= −) = .5	p(S_i= − | D= +) = .2	p(T_i= − | S_i= +) = .1
	p(S_i= + | D= −) = .2	p(T_i= + | S_i= −) = .1
	p(S_i= − | D= −) = .8	p(T_i= − | S_i= −) = .9

$$p(h|X_t) = \frac{p(X_t|h) \cdot p(h)}{p(X_t)}. \quad (5)$$

The likelihood is determined by the causal graph under consideration,

$$p(X_t|h) = p(T_1,...,T_n|h) = \prod_{i=1}^{n} p(T_i|\text{parents}(T_i),h) \quad (6)$$

where n represents the number of cues (or tests) and T_i is a random variable representing the result on test i. To model learning, we define the posterior of one trial as the prior for the next trial where the initial prior is $1/|H|$ where $|H|$ is simply the number of elements or graphs in the hypothesis space. Thus, the probability of the disease being present on trial t is given by

$$p(D=+|X_t) = \sum_{h \in H} p(D=+|X_t,h) \cdot p(h) \quad (7)$$

where $p(D=+|X_t,h)$ is easily calculated from the causal graph h. Please note that in the current formulation, only $p(h)$ is learned and not $p(T_i|S_j)$ or $p(S_i|D)$.

The Information Aggregation Stage Our approach towards information aggregation differs from the one taken by Budescu and Yu. Specifically, we assume that decision makers first attempt to reconstruct the data observed by advisers before calculating the probability of a target event (e.g. the presence of the disease). In this way, decision makers use both the advisers' probability estimates along with information about which cues the advisers saw, not merely how many cues the advisers saw. In the Bayesian model presented by Budescu and Yu, cue information was not inferred in the decision maker's aggregation process.

Suppose that there are J advisers reporting the probability of the presence of the disease to a decision maker. Further, assume that each adviser sees the results of a subset of all possible binary tests (or cues). The decision maker knows which tests each adviser saw, but not the results of the tests. Let p_j be the probability the disease is present reported by adviser j. (The probability the disease is absent is $1 - p_j$.) The decision maker then infers a belief distribution over the possible data that adviser j might have observed. Specifically, we define the data space as the set of all possible complete data vectors. Thus, for n cues, the data space has 2^n elements since the cues are assumed to be binary. In our experiment, $n = 3$ and the data space has $2^3 = 8$ elements. For simplicity, let us index the elements of the data space as X_i where $i \in \{1,2,3,...,2^n\}$. Then, for adviser j we have that the posterior probability of data vector X_i is

$$p(X_i|p_j,h) = \frac{p(p_j|X_i,h) \cdot p(X_i|h)}{p(p_j|h)} \tag{8}$$

where $p(p_j|h) = \sum_{i=1}^{2^n} p(p_j|X_i,h) \cdot p(X_i|h)$ and where we define $p(p_j|X_i,h) = 1$ if $p(X_i|h) = p_j$ and 0 otherwise. By examining the relationship between probabilities, $p(X_i|h)$, from the graphical structure and probabilities, p_j, from the adviser, we assume that the adviser only produces probabilities in accord with the graphical structure. By restricting $p(p_j|X_i,h)$ to the set $\{0,1\}$, we also assume there is no noise in the adviser's stated value. The model can be altered to relax these assumptions, but we felt that this was not necessary for modeling the current experiment. We acknowledge that the lack of variability or bias in advisers' ratings in unrealistic, but the assumption that advisers produce probabilities in accordance with the graphical structure is parsimonious. In the case where a judge sees a partial set of cues X_j^* such as $\{+,?\}$, we define $p(p_j|X_i,h) = 1$ if $X_i \subseteq X_j^*$ and $p(X_i|h) = p_j$. For example, in the case with two tests, suppose adviser j sees a positive result on the first test and nothing on the second. In other words, probability p_j is based on the data $X_j^* = \{+,?\}$. Since we only want to define a probability distribution over complete data, we consider both $\{+,+\}$ and $\{+,-\}$ as candidate data vectors. By summing over all possible cue relationships, we have

$$p(X_i|p_j) = \sum_{h \in H} p(X_i|p_j,h) \cdot p(h). \tag{9}$$

Next, the decision maker creates a belief distribution over all possible data by combining across advisers:

$$p(X_i|\bigcap_j p_j) = \frac{\prod_j p(X_i|p_j)/p(X_i)^{J-1}}{\sum_i \prod_j p(X_i|p_j)/p(X_i)^{J-1}} \tag{10}$$

where $p(X_i) = \sum_{h \in H} p(X_i|h) \cdot p(h)$. This formulation is identical to the one Budescu and Yu used to calculate the posterior probability of the target event. However, we assume that the aggregation process occurs when decision makers reconstruct the possible data seen by advisers instead of when they evaluate the probability of a target event. Like Budescu and Yu, we assume that the advisers' forecasts, p_j, are conditionally independent.

Finally, the decision maker creates a Bayesian posterior using his or her own prior for cue correlations, $p(h)$:

$$p(D=+|X_i) = \sum_{h \in H} p(D=+|X_i,h)^* \cdot p(h). \tag{11}$$

To allow for the fact that decision makers might not learn the exact probabilities in the graphical models, we make the following adjustment

$$p(D=+|X_i,h)^* = \frac{p(D=+|X_i,h)^\gamma}{p(D=+|X_i,h)^\gamma + (1-p(D=+|X_i,h))^\gamma} \tag{12}$$

where $p(D=+|X_i,h)$ is calculated directly from the graphical model h, and γ is allowed to freely vary. Thus, the probability the disease is present is given by

$$p(D=+) = \sum_{i=1}^{2^n} p(D=+|X_i) \cdot p(X_i) \tag{13}$$

where $p(X_i) = p(X_i|\bigcap_j p_i)$.

Model Comparisons

Using the data collected in the experiment described above, we evaluated the three models of interest: the WMLO model, the naively adjusted Bayesian model, and our new Bayesian inference model. Since the WMLO model and the naively adjusted Bayesian model do not contain learning models, we only used the information aggregation data in fitting the models and performing model comparisons. Because there was no difference is aggregation performance for correlated versus uncorrelated tests, we fit the models to all 57 participants. We evaluated WMLO as a parameter free model and fit the naively adjusted Bayesian model and our new Bayesian model by minimizing the sum of the squared error (SSE) between the model predictions and the probability judgment data from the experiment. Future work may use Bayesian model comparison, but for current purposes the simpler point estimates suffice to demonstrate qualitative differences between models. Both the naively adjusted Bayesian model and our new Bayesian model contain one parameter (λ and γ respectively). We fit the three models to the individual data for the 44 information aggregation trials. Thus, we obtained three model fits for each subject for 57 subjects. For each subject, we compared the one parameter naively adjusted Bayesian model with the one parameter cue imputation Bayesian model and the parameter free weighted mean log-odds model. The mean squared error (MSE) averaged over subjects for each model is given in Table 3. From the table, we see that our new Bayesian model has the smallest MSE.

We also computed a quantitative goodness-of-fit measure to a qualitative partitioning of the data. For each subject,

Table 3: MSE for Three Aggregation Models

Model	MSE
WMLO	0.0182
Naive Adj Bayes	0.0700
Imputation Bayes	0.0149

we found the observed frequencies of responses in the four categories (extremify, match, compromise and other) given the normatively correct response. In other words, we computed a table similar to Table 1 except the proportions were replaced by frequencies. We then computed the same frequencies for the three models using the best fit parameters to each subject's data. We calculated the SSE between the observed frequencies and model frequencies for each subject. This procedure is very similar to computing a chi-square test-statistic except that we do not weight the differences between observed frequencies and model frequencies by the reciprocal of the model frequencies. We felt that using such weights placed too much emphasis on instances where models have low frequencies, and we have no theoretical conviction that these instances should be weighted heavily. Table 4 shows several descriptive statistics for the SSE values. From the table, we see that our new Bayesian model out performs the other two models.

Table 4: Frequency SSE for Three Aggregation Models

model	mean	median	std	min	max
WMLO	347	254	278	26	1154
Naive Adj Bayes	679	650	398	24	1512
Imputation Bayes	298	230	217	10	1014

We used a qualitative evaluation of the models to see how closely the models were capturing ordinal trends in the data. We calculated the Kendall rank correlation coefficient (or Kendall's τ) for every subject to measure the association between the subject's response proportions and the predicated proportions from the models in the four categories (extremify, match, compromise and other) given the normatively correct response. Like the previous comparison, we used the best fit model parameters for each subject. Table 5 shows descriptive statistics for the τ values. A τ value of 1 implies perfect agreement between the rankings and a τ value of -1 implies perfect disagreement between the rankings. From the table, we again see that our new Bayesian model performs better than the other models.

Discussion

The cue imputation Bayesian model gives a more comprehensive account of information aggregation than previously proposed models. Whereas the naively adjusted Bayesian model cannot account for compromising judgments and the WMLO model cannot account for extremifying judgments, the new Bayesian model can account for both. The model provides a rational explanation of information aggregation by assuming that decision makers use the probability estimates of advisers to infer the underlying data before calculating the probability of an event. Thus, the model postulates that decision makers use all of the available information in an environment. Decision makers do not merely use the number of cues advisers saw, but they use this knowledge along with probability estimates from the advisers to infer information. Often, the information inferred by decision makers is more informative than the information seen by either adviser.

Table 5: Rank Correlation for Three Aggregation Models

model	mean	median	std	min	max
WMLO	0.24	0.26	0.22	-0.32	0.64
Naive Adj Bayes	0.10	0.09	0.21	-0.29	0.53
Imputation Bayes	0.36	0.38	0.17	-0.09	0.64

Our goal was to illustrate that it is possible to provide an account of probability judgments in situations of information aggregation by using the axiomatic principles of probability theory rather than heuristics. While heuristic models such as WMLO provide an initial means for studying probability judgment phenomena, these approaches lack a rational foundation and only provide an ad hoc explanation for probability judgment phenomena. However, probabilistic models such as the Bayesian model introduced in this paper offer a more coherent account of human probability judgments.

Acknowledgments

The first author was supported by the NSF/IGERT Training Program in the Dynamics of Brain-Body-Environment Systems at Indiana University.

References

Bordley, R. (1982). A multiplicative formula for aggregating probability assessments. *Management Sci*, *28*, 1137-1148.

Budescu, D., & Yu, H.-T. (2006). To bayes or not to bayes? a comparison of two classes of models of information aggregation. *Decision Analysis*, *3*, 145-162.

Budescu, D., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *J. Behav. Decision Making*, *20*, 153-177.

Genest, C., & Schervish, M. J. (1985). Modeling expert judgements for bayesian updating. *Ann. Statist*, *13*, 1198-1212.

Karmarkar, U. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organ. Behav. Human Performance*, *21*, 61-72.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press / Elsevier.