People’s conditional probability judgments follow probability theory (plus noise)

Fintan Costello,[1*] Paul Watts[2]

[1]School of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, Ireland
[2]Department of Mathematical Physics
National University of Ireland, Maynooth, Co Kildare, Ireland

[*]To whom correspondence should be addressed; E-mail: fintan.costello@ucd.ie.

Running head: CONDITIONAL PROBABILITY

Abstract

A standard view in current psychology is that people estimate probabilities using various 'heuristics' or rules of thumb that do not follow the normative rules of probability theory. We present a model where people estimate conditional probabilities such as $P(A|B)$ (the probability of A given that B has occurred) by a process that implements standard frequentist probability theory but is subject to random noise. This model accounts for various results from previous studies of conditional probability judgment. This model predicts that people's conditional probability judgments will agree with a series of fundamental identities in probability theory whose form cancels the effect of noise, while deviating from probability theory in other expressions whose form does not allow such cancellation. Two experiments that strongly confirm these predictions, with people's estimates agreeing closely with probability theory for the noise-cancelling identities, but deviating from probability theory (in just the way predicted by the model) for other identities. This new model subsumes an earlier model of unconditional or 'direct' probability judgment which explains a number of systematic biases seen in direct probability judgment (Costello and Watts, 2014). This model may thus provide a fully general account of the mechanisms by which people estimate probabilities.

keywords: probability; conditional probability; rationality; biases; heuristics

1 Introduction

A *conditional probability* $P(A|B)$ represents the chance that some event A will occur, given that some event B has definitely occurred. People estimate and use conditional probabilities very frequently in everyday life (for example, when I see dark clouds on the horizon and conclude, given those clouds, that rain is likely later). These probabilities are also central to critical decision making (for example, when a lawyer estimates the chances of winning or losing a case given a piece of evidence, and so decides whether or not to proceed to trial). Indeed, conditional probabilities play a fundamental role in many aspects of learning, reasoning, inference, and decision making under uncertainty. But how do people estimate the probability $P(A|B)$, given their knowledge about A and B? What mental processes underlie people's estimation of conditional probabilities?

Researchers have examined people's conditional probability judgment in various different ways. Perhaps the best-known approach involves presenting people with a kind of mathematical problem where they are given numerical values for the probabilities $P(A)$, $P(B)$ and $P(B|A)$ and then asked to estimate the conditional probability $P(A|B)$ (with their answers compared with the normatively correct value from probability theory). Well known examples of this approach are Eddy's 'breast cancer' problem (described in Gigerenzer and Hoffrage, 1995) and Kahneman and Tversky's 'taxi-cab' problem (Kahneman and Tversky, 1982). These studies reveal various reliable errors and biases in people's manipulation of presented probabilities: people tend to erroneously neglect the base rate $P(A)$, and have a tendency to confuse the conditional probabilities $P(A|B)$ and $P(B|A)$ (the 'inverse fallacy'). However, because these studies ask people to estimate one conditional probability $P(A|B)$ in terms of another conditional probability $P(B|A)$ (leaving the source of $P(B|A)$ unexplained) they tell us little about how conditional probability judgments arise from people's knowledge of events A and B.

These ‘mathematical problem’ studies suggest that people are extremely poor at assessing conditional probabilities. This fits with a currently dominant view of people’s probabilistic reasoning, which is that

> In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors (Kahneman and Tversky, 1973, p. 237)

This ‘heuristics and biases’ view has a level of popularity rarely seen in psychology (with Kahneman recieving a Nobel Prize in part for this work) and has had a major impact in a number of areas (Gigerenzer and Gaissmaier, 2011, Shafir and Leboeuf, 2002, Ariely, 2009, Kahneman, 2011, Camerer et al., 2003, Sunstein, 2000, Eva and Norman, 2005, Williams, 2010, Hicks and Kluemper, 2011, Bondt and Thaler, 2012). Studies which directly investigate people’s conditional probability judgment for events they have experienced, however, report results that in many ways contradict these findings. In these studies, people are not given a set of probabilities $P(A), P(B)$ and $P(B|A)$ and asked to estimate the conditional probability $P(A|B)$ from those values; instead people are simply asked to estimate probabilities such as $P(A), P(B), P(B|A)$ and $P(A|B)$ from their own experience with the events in question. These studies show low rates of occurrence of base-rate neglect and the ‘inverse fallacy’ (for a detailed review of these results see Koehler (1996); for a more general discussion of this ‘description-experience gap’, see Hertwig and Erev (2009)). More recent studies also suggest that people’s conditional probability estimates can closely follow the normative rules of probability theory in some ways while deviating from those rules in others. For example, in a recent study Fisher and Wolfe (2014) found that the addition form of Bayes' rule

$$P(A|B)P(B) - P(B|A)P(A) \;=\; 0$$

held reliably in people's probability estimates, just as required by probability theory. Zhao et al. (2009), by contrast, found that the requirement

$$P(A|B) \;=\; \frac{P(A \wedge B)}{P(B)}$$

(which defines conditional probability in standard probability theory) was reliably violated in people's probability estimates, contrary to the rules of probability theory. Since the addition form of Bayes' rule follows directly from this definition of conditional probability, this presents an interesting and surprising conflict.

In this paper we describe a mathematical model that aims to explain how people estimate conditional probabilities. This model makes only two assumptions. First, it assumes that people estimate conditional probabilites by simply counting event occurrence as in standard frequentist probability theory (in the model, estimating $P(A|B)$ involves counting how often event A occurs in the set of instances of event B). Second, it assumes that this counting process is subject to symmetric random error or noise, and so sometimes an instance of event A can be mistakenly counted as *not A* ($\neg A$) and sometimes an instance of $\neg A$ can mistakenly counted as A. With just these two assumptions, the model can account for the above results on people's conditional probability estimation and Bayes rule. More importantly, this model allows us to make a number of novel predictions. Specifically, the model identifies a number of probabilistic expressions (or *identities*) where the effects of random noise should cancel out. The model thus predicts that people's probability estimates should agree with probability theory for all these identities (showing no systematic bias away from the requirements of probability theory for those identities). For other identities, the model predicts systematic bias away from the requirements of probability theory, with the same amount of deviation for each identity. In this paper we describe two experiments testing these predictions. The results support our model: for these noise-cancelling identities, people's probability estimates agreed with probability theory, showing no systematic bias away from the values required by probability theory for those identities.[1] For the other identities, however, people's estimates systematically deviated from the requirements of probability theory in just the way predicted by the model. Taken together, these results support the view that people reason using a mechanism that follows probability theory, with bias in people's probability estimates being caused by random variation or noise in the reasoning process.

The model we present here goes significantly beyond an earlier model of people's estimates for direct or 'marginal' probabilities such as $P(A)$ (Costello and Watts, 2014). That earlier model, based on the same two assumptions, explained a number of systematic biases seen in people's direct probability estimates (biases such as conservatism, subadditivity, and the conjunction and disjunction fallacy). Because that model addressed only direct probabilities, however, it was of relatively narrow applicability: most everyday probability judgments involve conditional probabilities (the chance of rain, given clouds; the chance of winning a legal case, given the evidence; the chance of successful treatment for a disease, given a patient's symptoms; the chance of an experimental hypothesis being correct, given the data). The model we present here covers all such judgments, and indeed applies equally to conditional and direct probability estimation. (A direct probability $P(A)$ is equivalent to a conditional probability $P(A|B)$ in the special case where $P(B) = 1$. Our model of conditional probabilities exactly reduces to the earlier model of direct probabilities when $P(B) = 1$, and so inherits that model's account of biases in such probabilities). This new model can thus be seen as a fully general account of all

[1] To be clear: when we say that people's probability estimates 'agree' with probability theory in these identities, we do not mean that every single individual estimate exactly matched the value required by probability for that identity. Instead we take the standard view in experimental science, which assumes random error in measurement and sees experimental results as agreeing with a predicted value when individual measurements have a distribution that is peaked at, and distributed symmetrically around, that predicted value.

forms of probability estimation.

The paper is organised as follows. In the first section we set the scene by briefly presenting our model of direct probability estimation, and describe its predictions of systematic bias for some expressions and no bias for others. In the second section we present our new model of estimation for both conditional and direct probabilities, and describe similar, though more general, predictions. In the third section we show that this model is consistent with some previous results on people's conditional probability estimation. In the fourth section we describe an experiment testing, and confirming, the model's predictions for a set of frequently seen events. In the fifth section we describe a second experiment testing and confirming these predictions for a set of unique events. In the final section we draw some general conclusions from this work.

2 Our model of probability estimation

Our model assumes that people's probability judgments are produced by a mechanism that is fundamentally rational, but is perturbed in various ways by the systematic effects or biases caused by purely random noise or error. Here we are following a line of research leading back at least to Thurstone (1927) and continued by various more recent researchers (see, e.g. Dougherty et al., 1999, Erev et al., 1994, Hilbert, 2012). Our contribution here is to extend this work to give a general account for both conditional and direct probabilities, and to derive various expressions that 'cancel out' systematic biases due to random noise and so demonstrate agreement with probability theory in people's probability estimates.

Our model assumes a memory that returns or produces a set of items representing individual episodes or events. We take $P(A)$ to represent the 'true' probability of event A (that is, the proportion of items in memory that represent A). We take $P_E(A)$ to represent an individual estimate of the probability of event A, and take $\langle P_E(A) \rangle$ to represent the expectation value or mean of these estimates for A: this is the value we would expect to get if we averaged an infinite number of individual estimates for $P_E(A)$.

In standard probability theory, the probability of some event A is estimated by drawing a random sample of events, counting the number of those events that are instances of A, and dividing by the sample size. The expected value of these estimates is $P(A)$, the probability of A. We assume that people estimate the probability of some event A in exactly this way: randomly sampling events from memory, counting the number that are instances of A, and dividing by the sample size. If this counting process was error-free, people's estimates would have an expected value of $P(A)$. Human memory is subject to various forms of random error, however. To reflect this we assume events have some chance $d < 0.5$ of randomly being counted incorrectly: there is a chance d that a $\neg A$ (*not A*) event will be incorrectly counted as A, and the same chance d that an A event will be incorrectly counted as $\neg A$. A randomly sampled event will be counted as A if the event truly is A and is counted correctly (with a probability $(1-d)P(A)$, since $P(A)$ events are truly A and events have a $1-d$ chance of being counted correctly), or if the event is truly $\neg A$ and is counted incorrectly as A (with a probability $(1-P(A))d$, since $1-P(A)$ events are truly $\neg A$, and events have a d chance of being counted incorrectly). The expected value for a noisy estimate for the probability of A is thus

$$\langle P_E(A) \rangle \quad = \quad (1-d)P(A) + (1-P(A))d = (1-2d)P(A) + d \tag{1}$$

with individual estimates varying independently around this expected value.[2] This expected value for estimates embodies a regression towards the center, due to random noise: estimates are systematically biased away from the 'true' probability $P(A)$, such that estimates will tend to be greater than $P(A)$ when $P(A) < 0.5$, and will tend to be less than $P(A)$ when $P(A) > 0.5$, and will tend to equal $P(A)$ when $P(A) = 0.5$. In previous work (Costello and Watts,

[2] Since the model assumes that an individual estimate $P_E(A)$ is produced by drawing a random sample of N events, counting the number that are read as A, and dividing by the sample size, the distribution of estimates around their expected value follows the binomial proportion distribution $B(p, N)/N$ with $p = (1-2d)P(A) + d$.

2014) we showed that this pattern of regression explains systematic biases in people's probability judgments such as underconfidence (people's tendency to overestimate probability for low-probability events and understimate probability for high probability events) and subadditivity (people's tendency to give a probability estimate for a disjunction of exclusive events $P_E(A_1 \vee A_2 \ldots A_n)$ that is less than the sum of their probability estimates for the individual events $P_E(A_1) + P_E(A_2) + \ldots + P_E(A_n)$).

While this model predicts these systematic biases in people's judgments, it also predicts unbiased agreement with probability theory when people's judgments are combined in various probabilistic expressions. In particular, the model predicts that, on average, people's probability estimates will match the requirements of probability theory for the identities

$$\begin{aligned} P(A) + P(B) - P(A \wedge B) - P(A \vee B) &= 0 \\ P(A) + P(B \wedge \neg A) - P(B) - P(A \wedge \neg B) &= 0 \end{aligned}$$

(the first two identities in Table 1). Probability theory requires that when the terms in these identities are summed, the resulting value must be 0 for all events A and B. Our model also predicts that, on average, people's estimates for these expressions will also sum to 0. More strictly, our model predicts that the expected value for these identities - that is, the average of an infinite number of values of these identities - will exactly equal 0. For the average of some finite sample of values for these identities, the prediction is a value close to and varying randomly around 0, approaching 0 more and more closely as the sample size increases. For example, consider the first identity (probability theory's 'addition law'). Suppose we ask people to estimate $P(A), P(B), P(A \wedge B)$ and $P(A \vee B)$ and combine each person's estimates in the form of the addition law. Since the expected value of a sum is equal to the sum of expected values of its terms, the expected value for this combination is, using Equation 1,

$$\langle P_E(A)\rangle + \langle P_E(B)\rangle - \langle P_E(A \wedge B)\rangle - \langle P_E(A \vee B)\rangle = \\ (1-2d)\,[P(A)+P(B)-P(A \wedge B)-P(A \vee B)] + 2d - 2d = 0$$

and so the expected value for the addition law in people's estimates will be 0 just as required in probability theory. Since individual values for this sum are perturbed by random noise in the individual estimates, we expect these individual values to be distributed symmetrically around that mean of 0. We also expect the average for some finite sample of values for this identity to be close to 0 (varying randomly around 0, and approaching 0 more and more closely as the sample size increases). The same result holds for the second identity. A number of experiments have shown that these identities do in fact hold in people's probability judgments, just as predicted by the model: when we ask people to estimate probabilities for the terms in these identities for a range of events, and then combine each person's estimates according to the identity, the values obtained are distributed approximately symmetrically around a mean of 0, as required by probability theory and predicted by our model (Costello and Watts, 2014, Costello and Mathison, 2014, Fisher and Wolfe, 2014).

This model also predicts that people's probability estimates will violate the requirements of probability theory for the first 4 identities in Table 2, with the same degree of violation for each identity. Probability theory requires that these identities must also sum to 0 for all events A and B. Substituting our model's expression for the expected value for estimates of each term (just as in the 'addition law' case) gives an overall positive expected value of d, violating the requirement of probability theory. For example, the estimated value of the expression in Identity 9 is

$$\langle P_E(A)\rangle + \langle P_E(B \wedge \neg A)\rangle - \langle P_E(A \vee B)\rangle = \\ (1-2d)\,[P(A)+P(B \wedge \neg A)-P(A \vee B)] + 2d - d = d$$

Table 1: Probability theory identities that our model predicts will hold in people's probability estimates. Our model predicts if these identities are computed from people's individual probability estimates for any pair of events A, B, the values obtained will be symmetically distributed around a mean of 0, the value required by probability theory. Equivalently, our model predicts that the positive and negative terms in these identities should, on average, be equal in people's probability judgments.

Identity	positive terms		negative terms		predicted value
1	$(P(A) + P(B))$	$-$	$(P(A \wedge B) + P(A \vee B))$	$=$	0
2	$(P(A) + P(B \wedge \neg A))$	$-$	$(P(B) + P(A \wedge \neg B))$	$=$	0
3	$P(A\|B)P(B)$	$-$	$P(B\|A)P(A)$	=	0
4	$(P(A\|B)P(B) + P(A\|\neg B))$	$-$	$(P(A\|\neg B)P(B) + P(A))$	=	0
5	$(P(B\|A)P(A) + P(B\|\neg A))$	$-$	$(P(B\|\neg A)P(A) + P(B))$	=	0
6	$(P(B\|A)P(A) + P(A\|\neg B))$	$-$	$(P(A\|\neg B)P(B) + P(A))$	=	0
7	$(P(A\|B)P(B) + P(B\|\neg A))$	$-$	$(P(B\|\neg A)P(A) + P(B))$	=	0
8	$(P(A\|\neg B) + P(B) + P(B\|\neg A)P(A))$	$-$	$(P(B\|\neg A) + P(A) + P(A\|\neg B)P(B))$	=	0

Again, a number of experiments have shown that these identities are indeed violated in people's probability estimates, in just the way predicted by the model.

Note that the parameter d in this model gives a 'first-order approximation' to the effect of random error in probability estimation: it represents the assumption that there is random noise in probability estimation, and that noise is, to a first approximation, the same for all types of events. A more detailed version of this model goes beyond this first approximation by addressing the various factors that could influence random error in memory. Among the factors that we expect will influence random error rates are individual differences among participants, with some people having a lower propensity to random error than others (we saw just this pattern in our previous work (Costello and Watts, 2014)); task type, with some tasks being associated with higher rates of random error than others; and event type, with probability estimates for more complex events (such as conjunctions and disjunctions) being associated with slightly higher rates of random error.

In Costello and Watts (2016a,b) we described a more detailed second-order version of our model which assumed error rates of $d + \Delta d$ for conjunctions and disjunctions versus rates of

Table 2: Probability theory identities that our model predicts will be violated in people's estimates. Our model predicts if these identities computed from people's individual probability estimates for any pair of events A, B, the values obtained will be positive, contrary to the requirements of probability theory, and that identities $9 \ldots 12$ will have approximately the same value (equal to d) and identities $13 \ldots 16$ will have approximately half that value (equal to $d/2$).

Identity	probability theory identity		predicted value in model
9	$P(A) + P(B \wedge \neg A) - P(A \vee B)$	$= 0$	d
10	$P(B) + P(A \wedge \neg B) - P(A \vee B)$	$= 0$	d
11	$P(A \wedge \neg B) + P(A \wedge B) - P(A)$	$= 0$	d
12	$P(B \wedge \neg A) + P(A \wedge B) - P(B)$	$= 0$	d
13	$P(A \wedge B) - P(A\|B)P(B)$	$= 0$	$d/2$
14	$P(A \wedge B) - P(B\|A)P(A)$	$= 0$	$d/2$
15	$P(A \wedge B) - P(A) + P(A\|\neg B)(1 - P(B))$	$= 0$	$d/2$
16	$P(A \wedge B) - P(B) + P(B\|\neg A)(1 - P(A))$	$= 0$	$d/2$

d for single events, where Δd is represents a small adjustment in error rates for more complex conjunctive and disjunctive events. This assumption follows the standard statistical concept of propagation of error, which states that if two variables A and B are subject to random error, then a complex variable (e.g. $A \wedge B$) that is a function of those two variables will have a higher rate of error than either variable on its own. In this second-order model we thus have an expected value for estimates of $P(A \wedge B)$ and $P(A \vee B)$ of

$$\begin{aligned} \langle P_E(A \wedge B) \rangle &= (1 - 2[d + \Delta d])P(A \wedge B) + [d + \Delta d] \\ &= (1 - 2d)P(A \wedge B) + d + \Delta d(1 - 2P(A \wedge B)) \end{aligned}$$

$$\begin{aligned} \langle P_E(A \vee B) \rangle &= (1 - 2[d + \Delta d])P(A \vee B) + [d + \Delta d] \\ &= (1 - 2d)P(A \vee B) + d + \Delta d(1 - 2P(A \vee B)) \end{aligned}$$

These expected values follow the values that would be given by Equation 1 for $A \wedge B$ and $A \vee B$: they deviate from those values by a second-order correction term bounded by $\pm \Delta d$, a value that we assume is small.

This more detailed model gives an accurate account for the occurrence of the conjunction fallacy in people's probability estimates. This fallacy occurs when people give a conjunctive probability estimate $P_E(A \wedge B)$ that is higher than their estimate for a constituent of that conjunction, $P_E(A)$, contrary to the requirements of probability theory. Experimental results show that the conjunction fallacy can occur at high rates for some conjunctions, and much lower rates for others (Fisk and Pidgeon, 1996). In this model, higher noise rates for conjunctions than constituents ($d + \Delta d$ versus d) will cause greater regression towards the center for conjunctive probabilities than for constituent probabilities, producing, in some cases, average estimates where $\langle P_E(A \wedge B)\rangle > \langle P_E(A)\rangle$ and so causing high conjunction fallacy rates.

We applied a simulation implementing this model to Fisk and Pidgeon's experimental results on the rate of occurrence of the conjunction fallacy in people's probability judgments (Costello and Watts, 2016a). With a very low value of Δd ($= 0.015$) the simulation produced results that agreed with Fisk and Pidgeon's results (mean absolute difference between observed and simulated fallacy rates of 3.9%; mean absolute difference between observed and simulated probability estimates of 0.02; correlation between observed and simulated fallacy rates of $r = 0.95, p < 0.0001$; correlation between observed and simulated probability estimates of $r = 0.96, p < 0.0001$). This agreement held for both high and low fallacy rates: for example, 69.2% of participants produces a fallacy for one conjunction in Fisk and Pidgeon's data, and the simulation generated a fallacy rate of 68.8% for that conjunction; only 6.6% of participants produced a fallacy for another conjunction in Fisk and Pidgeon's data, and the simulation generated a fallacy rate of 5.3% for that conjunction. The simulation also accounted for the observed agreement with probability theory for noise-cancelling identities such as the addition law (for details, see Costello and Watts, 2016a).

From the equations above, the terms in our first-order approximation dominate the expected-value predictions in our model: expected-value predictions from the more complex second-order model closely follow those of the basic first-order model, varying around those predictions with a range proportional to the small correction term Δd. For example, where our first-order model predicts a expected value of 0 for Identity 1, by substituting the above equations we see that this second-order model predicts an expected value of

$$\langle P_E(A)\rangle + \langle P_E(B)\rangle - \langle P_E(A \wedge B)\rangle - \langle P_E(A \vee B)\rangle = 2\Delta d[P(A \wedge B) + P(A \vee B) - 1]$$

Since $-1 \leq [P(A \wedge B) + P(A \vee B) - 1 \leq 1$ necessarily holds, this second-order model predicts a expected value for the identity that is within $-2\Delta d$ and $+2\Delta d$ of 0; or taking the value $\Delta d = 0.015$ from our simulation, within -0.03 and 0.03 of 0. The same prediction holds for Identity 2. Given that the first-order model in any case expects the average for some finite sample of values for this identity to vary randomly around 0 (due to random variance in samples), there is little difference between the predictions of this second order model, and the predictions of the simpler first-order model, for values of these identities. Since our focus here is not on the conjunction fallacy, but on simple conditional probabilities such as $P(A|B)$ (to which error rate adjustments such as Δd do not apply), in this paper we consider the first-order model only: for simplicity we take the error rate d as representing a constant rate of random error in memory for all events.

3 Estimating conditional probabilities

We assume that people estimate a conditional probability $P(A|B)$ by counting occurrences of A in a set of instances of B, just as in standard probability theory. A number of previous researchers have given related accounts of conditional probability estimation based on counting instances in samples (see, for example Fox and Levav, 2004, Fiedler et al., 2000). The main novelties in our account are that we provide a specific mechanism for the effects of random noise on this counting process; that our account covers both conditional probabilities and direct probabilities; and that our model makes specific predictions about the value of various conditional probability identities.

We assume that people estimate $P(A|B)$ by drawing a random sample of instances of B, counting the number that are also A, and dividing by the sample size. As before, we assume some chance of random error d in this counting process. Given this random error there are two mutually exclusive ways in an item can be read as an instance of event B: (i) when the item truly is an instance of B and is read correctly (this occurs with probability $(1-d)P(B)$); and (ii) when the item is actually $\neg B$ but is read incorrectly as B (this occurs with probability $d(1-P(B))$).

We first take case (i). Given that a randomly sampled item is read as B, the probability that the item is truly an instance of B is

$$\frac{(1-d)P(B)}{(1-d)P(B)+d(1-P(B))} = \frac{(1-d)P(B)}{(1-2d)P(B)+d}$$

with the denominator here representing the average number of items that will be read as B, and the numerator the average number of those that have been read correctly.

Given that we truly have an instance of B, there are two mutually exclusive ways in which that item can be read as A: when the item is indeed an instance of A and is read correctly, or when the item is actually $\neg A$ and is read incorrectly as A. Since $P(A|B)$ is the probability of an item being truly an instance of A given that it is truly an instance of B, the first possibility occurs with probability $(1-d)P(A|B)$; since $1-P(A|B)$ is the probability of an item being truly $\neg A$ given that it is truly an instance of B, the second possibility occurs with probability $d(1-P(A|B))$. The sum of these two probabilities is $(1-2d)P(A|B)+d$, and so the overall probability of an instance being read as A given that it is truly B and was correctly read as B is

$$\frac{(1-d)P(B)\left[(1-2d)P(A|B)+d\right]}{(1-2d)P(B)+d} \tag{2}$$

We next take case (ii). Given that a randomly sampled item is read as B, the probability that the item is truly an instance of $\neg B$ is

$$\frac{d(1-P(B))}{(1-2d)P(B)+d}$$

Again, given that we truly have an instance of $\neg B$, there are two mutually exclusive ways in which that item can be read as A: when the item is indeed an instance of A and is read correctly, or when the item is actually $\neg A$ and is read incorrectly. Reasoning as before, we see that the overall probability of an instance being read as A given that it is truly $\neg B$ but was read as B is

$$\frac{d(1-P(B))\,[(1-2d)P(A|\neg B)+d]}{(1-2d)P(B)+d} \tag{3}$$

Since (i) and (ii) are mutually exclusive and cover all possibilities, the sum of Equations 2 and 3 gives our predicted value for $\langle P_E(A|B)\rangle$, the average estimate for the conditional probability $P(A|B)$. Adding the two and using the probability theory identities

$$\begin{aligned} P(B)P(A|B) &= P(A\wedge B) \\ (1-P(B))P(A|\neg B) &= P(A\wedge\neg B) = P(A)-P(A\wedge B) \end{aligned}$$

we get

$$\langle P_E(A|B)\rangle = \frac{(1-2d)^2P(A\wedge B)+d(1-2d)\,[P(A)+P(B)]+d^2}{(1-2d)P(B)+d} \tag{4}$$

Just as with Equation 1, this average is systematically biased away from the 'true' probability $P(A|B)$, and subject to regression towards the center such that estimates will tend to be greater than $P(A|B)$ when $P(A|B)$ is low, and less than $P(A|B)$ when $P(A|B)$ is high. Where Equation 1 is regressive towards 0.5, Equation 4 is regressive towards a point that depends on the probabilities $P(A\vee B)$ and $P(A\wedge B)$. To see this, consider that the difference between the 'true' conditional probability $P(A|B)$ and the average estimated conditional probability, $\langle P_E(A|B)\rangle$, is equal to

$$P(A|B)-\langle P_E(A|B)\rangle = \frac{d[P(A|B)-[(1-2d)(P(A\vee B)-P(A\wedge B))+d]]}{(1-2d)P(B)+d}$$

(substituting Equation 4 and simplifying). Taking

$$R_{A,B} = (1-2d)(P(A \vee B) - P(A \wedge B)) + d$$

we see that $\langle P_E(A|B)\rangle < P(A|B)$ when $R_{A,B} < P(A|B)$, and $\langle P_E(A|B)\rangle > P(A|B)$ when $R_{A,B} > P(A|B)$, and so the average conditional probability estimate $\langle P_E(A|B)\rangle$ is regressive towards the point at which the conditional probability $P(A|B) = R_{A,B}$. As $P(B)$ approaches 1 the regression point (the point at which $P(A|B) = R_{A,B}$) approaches 0.5; when $P(B) = 1$ we have $P(A|B) = P(A)$ and $R_{A,B} = (1-2d)(1-P(A)) + d$ and so $P(A|B) = R_{A,B}$ when $P(A) = 0.5$, and regression is towards 0.5 as with Equation 1.

An important point to note here is that this model of conditional probabilities does not derive an estimate for $P(A|B)$ in terms of estimates for $P(A \wedge B)$ and $P(B)$. Instead, this model assumes that $P(A|B)$ is estimated by sampling instances of B, and counting the proportion of those that are instances of A. Since this model doesn't involve estimates for conjunctions like $P(A \wedge B)$, it doesn't involve the aforementioned error adjustment term Δd that applies to such conjunctions.[3]

Note that a direct probability $P(A)$ is, in probability theory, equivalent to a conditional probability $P(A|B)$ where the conditioning event B has a probability of 1 . Rearrangement shows that when $P(B) = 1$, Equation 4 reduces to Equation 1, our expression for direct probability estimation. Equation 4 thus completely describes all probability estimates, both direct and conditional, in this model. While Equation 4 appears complicated, it follows directly from two simple assumptions: that probabilities are estimated by counting event occurrence (in accordance with probability theory) and that this counting process is subject to random noise.

It is important to stress here that we are not suggesting that people explicitly use Equation 4 when estimating probabilities. Instead, our suggestion is that people estimate probabilities by

[3] More complex conditionals, such as $P(A \wedge B|C)$ or $P(A|B \wedge C)$, would involve this Δd term. Expected value expressions for such conditionals would be significantly more complex than the expected value for $P(A|B)$ in Equation 4. One aim for future work is to develop and test expressions for these more complex conditionals.

simply counting event occurrence (subject to random noise); Equation 4 describes the average estimate we would expect to see when people estimate probabilities using this noisy counting mechanism.

3.1 Predictions

From probability theory we have a number of identities whose value must be 0 for all events A and B. One such identity is the addition form of Bayes' rule (Identity 3 in Table 1). Our model predicts that this identity should also hold in people's probability judgments, on average. To see this, suppose we ask people to estimate $P(A), P(B), P(A|B)$ and $P(B|A)$ and for each person we take the products $P(A|B)P(A)$ and $P(B|A)P(A)$. Since estimates vary independently, the expected value of the products is equal to the product of the expected values of their constituents, giving

$$\begin{aligned}\langle P_E(A|B)P_E(B)\rangle &= \langle P_E(A|B)\rangle \langle P_E(B)\rangle \\ &= (1-2d)^2 P(A \wedge B) + d(1-2d)[P(A)+P(B)] + d^2\end{aligned}$$

and similarly

$$\begin{aligned}\langle P_E(B|A)P_E(A)\rangle &= \langle P_E(B|A)\rangle \langle P_E(A)\rangle \\ &= (1-2d)^2 P(A \wedge B) + d(1-2d)[P(A)+P(B)] + d^2\end{aligned}$$

and so

$$\langle P_E(A|B)P_E(B) - P_E(B|A)P_E(A)\rangle = \langle P_E(A|B)P_E(B)\rangle - \langle P_E(B|A)P_E(A)\rangle = 0$$

Thus our model predicts that the average value of this identity, computed from people's individual probability judgments, should equal 0 as required by probability theory (more strictly, the prediction is that the expected value for this identity - that is, the average of an infinite number of values of these identities - will exactly equal 0. For the average of some finite sample of values for this identity, the prediction is a value close to 0, approaching 0 more and more closely as the sample size increases). Since deviations from this expected average in individual estimates are due to random error in our model (which is equally likely to be positive or negative), we also expect that individual values for these identities will be approximately symmetrically distributed around 0.

Similar expansion and rearrangement gives the same result for Identities 4 through 8 in Table 1. Our model therefore predicts that these identities should all have an average value of 0 in people's estimates (matching the requirements of probability theory), and that individual values for these identities will be approximately symmetrically distributed around 0.

While this model predicts agreement with probability theory for the identities given above, it also predicts that Identities 13 through 16 in Table 2 should have a positive value in people's estimates, violating probability theory. For example, using the same substitutions as above we get an expected value for Identity 13 of

$$\begin{aligned}\langle P_E(A \wedge B) - P_E(A|B)P_E(B)\rangle = &\ (1-2d)P(A \wedge B) + d - (1-2d)^2 P(A \wedge B) \\ &- d(1-2d)[P(A) + P(B)] - d^2 \\ &= d(1-d) - d(1-2d)\left[P(A) + P(B) - 2P(A \wedge B)\right]\end{aligned}$$

Similar substitutions gives exactly the same expected value for identities $14, 15$ and 16. Probability theory requires that $0 \leq P(A) + P(B) - 2P(A \wedge B) \leq 1$ for all A and B, and since $d < 0.5$ by assumption, we see that

$$d^2 \leq \quad d(1-d) - d(1-2d)\left[P(A) + P(B) - 2P(A \wedge B)\right] \quad \leq d(1-d)$$

and values for this expression will be distributed between d^2 and $d(1-d)$ in a way that depends on $P(A) + P(B) - 2P(A \wedge B)$. The average value for $P(A) + P(B) - 2P(A \wedge B)$ (across uniformly distributed probabilities that are constrained to be consistent with probability theory) is 0.5, and so the average value for this expression is equal to $d/2$, the centerpoint of this range. Our prediction, therefore, is that Identities 13 through 16 should have, on average, a value of $d/2$, i.e. half the value of Identities 9 through 12.

4 Previous results on conditional probability estimation

Here we show that our model is consistent with the results of various previous studies involving conditional probability estimation: those of Fiedler et al. (2000), Zhao et al. (2009) and Fisher and Wolfe (2014).

Fiedler et al. (2000) carried out a series of studies examining the role of sampling in people's conditional probability estimation from experience. In these experiments participants first learned about the conditional relationship between events A and B by searching through a series of 'training cases' of those events (where A, represented, for example, having breast cancer, and B represented a positive mammography result; and the training cases represented a set of patient with positive or negative mammography results and with or without breast cancer). They were then asked to estimate the conditional probability $P(A|B)$. The same set of training cases was used for all participants; the objective conditional probabilities $P(A|B)$ were low (< 0.5) in these training cases. For one group of participants the training instances were organised according to values of the conditioning event or *predictor* B; for another group the training instances were organised by the conditioned event or *criterion* A. Participants in the predictor group gave conditional probability estimates that were close to the correct value; participants in the criterion group gave estimates that showed a significant bias away from that value, with estimates being systematically higher than the objective values. The general observation that participants in the predictor group gave results that were close to the correct value supports the idea that conditional probability estimates are produced by some form of sampling and counting as in our model. The more specific pattern of results is also broadly consistent with our model: in our model participants in the predictor group would find it easy to recall a set of instances of B (that is, would be less subject to random error) and so give a good estimate of the conditional probability $P(A|B)$, while participants in the criterion group would find it hard to recall a set of instances of B (that is, would be more subject to random error) and so would give a more biased estimate of the conditional probability. Further, since the objective conditional probabilities $P(A|B)$ were low, the model would predict a regressive bias causing estimates to be higher than objective values, just as observed.

Zhao et al. (2009) carried out two experiments where participants saw a range of different repeatedly-presented events (the occurrence of squares, circles and triangles of different colours). Participants saw 12 sets of these events, each set containing 20 individual coloured shapes (red squares, green triangles, and so on). The objectively correct probabilities for events in these sets were assigned by the experimenters at three levels, low, medium and high: there were 4 sets of each type. After seeing the 4 sets in a given level (and so seeing 80 individual shapes), participants were asked to estimate the probability $P(B)$ (e.g. probability a random shape is a square) on one presentation, the probability $P(A \wedge B)$ (e.g. probability a random shape is a red square) on another presentation, and the probability $P(A|B)$ (e.g. probability a random shape is red given that it is a square) on the third presentation. For each of these levels, the average of people's conditional probability estimates were relatively close to the objectively correct probability: all were within 0.07 of the correct probability in Experiment 1, and within 0.05 of the correct probability in Experiment 2. Both direct and conditional probability estimates tended to be higher than the correct probability for low probability events, and lower than the correct probability for high probability events: a pattern of regression towards the center, as predicted by our model (see Table 3). These patterns of relatively close agreement with the objective probability, and regression towards the center, are consistent with our model.

Table 3: Objective probability values and average probability estimates (SD) produced by participants in Experiments 1 and 2 of Zhao et al. (2009), along with estimates produced by a simulation implementing our model for the same objective probabilities (with $d = 0.05$). Note that in every case simulation estimates fell within one standard deviation of participant's estimates, both in Experiment 1 and Experiment 2.

Probability	objective value	Probability Estimates (Zhao et al., 2009)		Simulation
		Expt 1	Expt 2	
$P(B)$	0.3	0.31 (0.07)	0.29 (0.05)	0.32 (0.02)
$P(B)$	0.6	0.58 (0.07)	0.59 (0.08)	0.59 (0.02)
$P(B)$	0.9	0.86 (0.04)	0.85 (0.05)	0.86 (0.02)
$P(A \wedge B)$	0.1	0.19 (0.08)	0.17 (0.06)	0.15 (0.02)
$P(A \wedge B)$	0.4	0.5 (0.10)	0.51 (0.16)	0.42 (0.02)
$P(A \wedge B)$	0.8	0.8 (0.07)	0.78 (0.06)	0.77 (0.02)
$P(A\|B)$	0.33	0.35 (0.10)	0.33 (0.09)	0.36 (0.06)
$P(A\|B)$	0.67	0.63 (0.10)	0.63 (0.10)	0.65 (0.04)
$P(A\|B)$	0.89	0.82 (0.12)	0.83 (0.06)	0.84 (0.03)
$P(A \wedge B)/P(B)$	0.33	0.70 (0.32)	0.65 (0.25)	0.44 (0.08)
$P(A \wedge B)/P(B)$	0.67	0.93 (0.22)	1.00 (0.41)	0.71 (0.06)
$P(A \wedge B)/P(B)$	0.89	1.03 (0.35)	1.15 (0.68)	0.90 (0.04)

This model can also explain the apparent conflict between Zhao et al. (2009)'s results and those of Fisher and Wolfe (2014). Probability theory requires that

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \textit{ provided } P(B) > 0 \tag{5}$$

Zhao et al. (2009) found, however, that the conjunction ratio $P(A \wedge B)/P(B)$ did not predict $P(A|B)$ well. Instead they found that

$$\frac{P_E(A \wedge B)}{P_E(B)} - P_E(A|B) > 0$$

held reliably in people's probability estimates (this 'conjunction ratio' was reliably higher than people's conditional probability estimates). The difference between the conjunction ratio and people's conditional probability estimates was typically quite large (around 0.3, where probabilities range from 0 to 1), and was reliable across participants and across a range of probability values for A and B.

These results represent a systematic *bias* in people's probability judgments; that is, a systematic deviation from the requirements of probability theory for the ratio $P(A \wedge B)/P(B)$. By contrast, Fisher and Wolfe (2014) asked people to give estimates $P_E(A)$, $P_E(B)$, $P_E(A|B)$ and $P_E(B|A)$ for 34 different probability estimation problems. Each problem featured a short scenario followed by questions for $P(A)$,$P(B)$, $P(A|B)$, and $P(B|A)$. For example, one problem presented the scenario

> *Steve is 50 years old and has a sedentary lifestyle. He is a movie buff. When he comes home from his job as a computer programmer, he likes to watch movies from his movie collection and eat his favorite ice cream: double fudge, chocolate chip with sprinkles.*

and asked people to estimate the probabilites *P(Steve is obese)*, *P(Steve can do 50 push-ups)*, *P(Steve is obese* | *Steve can do 50 push-ups)* and *P(Steve can do 50 push-ups* | *Steve is obese).*

Problems varied in the probability of their constituents and in the conditional relationship between constituents, with some problems having a positive relationship (one constituent being more likely, given the other constituent) and others having a negative relationship (one constituent being less likely, given the other constituent, as in the Steve example). Participants saw each probability problem twice (at time 1 and time 2), allowing Fisher and Wolfe to approximately estimate the degree of random noise in participant's responses (by taking the absolute difference between estimates at those two times). They found that when people's individual estimates for these probability problems were combined in the addition form of Bayes' rule (identity 3 in Table 1). the values obtained were, on average, almost exactly equal to the value of 0 required by probability theory and predicted by our model: the average value for this identity in people's judgments was 0.008. (They also found that there was a reliable positive correlation, $r = .24$, $95\% CI[.21, .27]$, between degree of absolute deviation of values of this identity from 0 and a measure of noise obtained by summing the absolute differences in estimates at time 1 and time 2. Again, this is as predicted by our model, where higher levels of noise will cause a higher degree of variability in values of the identity, and so higher absolute deviation from 0).

As noted above, these two results present an interesting conflict. On the one hand, people's conditional probability estimates systematically violate probability theory's requirement for identity in Equation 5. On the other hand, people's conditional probability estimates on average agree closely with the addition form of Bayes' rule, which is a direct mathematical consequence of Equation 5. Our model, however, resolves this conflict. As we saw above, our model predicts an average value of 0 for the addition form of Bayes' rule, just as observed by Fisher and Wolfe (2014). Our model also predicts that, on average, the difference between the conjunction ratio and people's conditional probability estimates will be positive. From Equation

5 the expected value for the conjunction ratio is given by

$$\left\langle \frac{P_E(A \wedge B)}{P_E(B)} \right\rangle = \langle P_E(A \wedge B) \rangle \left\langle \frac{1}{P_E(B)} \right\rangle \textit{ provided } P_E(B) > 0$$

From Jensen's inequality we have

$$\left\langle \frac{1}{X} \right\rangle \geq \frac{1}{\langle X \rangle} \textit{ provided } X > 0$$

and so for the difference between the conjunction ratio and people's conditional probability estimates we have

$$\left\langle \frac{P_E(A \wedge B)}{P_E(B)} \right\rangle - \langle P_E(A|B) \rangle \geq \frac{\langle P_E(A \wedge B) \rangle}{\langle P_E(B) \rangle} - \langle P_E(A|B) \rangle$$

Rearranging the right-hand side of this inequality we get

$$\begin{aligned} \frac{\langle P_E(A \wedge B) \rangle}{\langle P_E(B) \rangle} - \langle P_E(A|B) \rangle &= \frac{(1-2d)P(A \wedge B) + d}{(1-2d)P(B) + d} \\ &\quad - \frac{(1-2d)^2 P(A \wedge B) + d(1-2d)[P(A) + P(B)] + d^2}{(1-2d)P(B) + d} \\ &= \frac{d[(1-2d)(2P(A \wedge B) - P(A) - P(B)) + 1 - d]}{(1-2d)P(B) + d} \end{aligned}$$

and we see this difference will be positive when

$$(1-2d)(2P(A \wedge B) - P(A) - P(B)) + 1 - d > 0$$

or equivalently, when

$$\frac{1-d}{1-2d} > P(A) + P(B) - 2P(A \wedge B) \tag{6}$$

Since d is positive (there is some chance of an error in recall) and by assumption $d < 0.5$ (a correct read from memory is more likely than an error), the left hand side of Equation 6 is always greater than 1. From probability theory, however, we have that

$$P(A) + P(B) - 2P(A \wedge B) = P(A \vee B) - P(A \wedge B)$$

and so the right hand side of Equation 6 is always less than or equal to 1. Our model therefore predicts that this difference will be positive and thus predicts that, on average, the conjunction ratio will be greater than the conditional probability, just as observed by Zhao et al. (2009).

We tested this account in more detail by applying a computer program that simulated the model to the experimental data from Zhao et al. (2009). This program took as input three probabilities $P_I(A)$, $P_I(B)$, and $P_I(A \wedge B)$, equal to the objective probabilities assigned by Zhao et al. (2009) in for each probability type. The program constructed a 'memory' containing 80 items (the same number of items seen by participants before estimating each probability), each item containing flags A, B, $A \wedge B$ indicating whether that item was an example of the given event. The occurrence of those flags in memory exactly matched the probabilities of the given event as specified by the three input probabilities. This program also took as input a noise parameter value d, representing the noise rate. To produce an estimate for the probability of some event B or $A \wedge B$, the program simply went through the memory reading the values of flags for that event and returned the proportion of flags that were read as true as its estimate for the probability of the given event. When reading flag values from memory to generate some probability estimate, the program was designed to have a random chance d of returning the incorrect value for a flag. To produce an estimate for some conditional probability $P(A|B)$, the program went through the memory identifying items whose flag for B was set (subject to random error in reading), counted the number of those items whose flag for A was also set (again subject to random error) and returned the proportion.

For a given set of objective input probabilities $P_I(A)$, $P_I(B)$, and $P_I(A \wedge B)$ from Zhao et al. (2009), each run of this simulation program generated a single noisy estimate for each of the probabilities $P_E(B)$, $P_E(A \wedge B)$ and $P_E(A|B)$. These represent a single individual's estimates for those probabilities. The program also generated an individual noisy estimate for the conjunction ratio $P(A \wedge B)/P(B)$, by simply dividing the individual estimates obtained for those probabilities. To apply this simulation to the data from Zhao et al. (2009) we ran this program $10,000$ times for each of the three objective probability levels low, medium and high, to obtain overall averages for these estimates. We fixed the noise rate d at 0.05 in all these runs, because this was the best-fitting value of d in a previous simulation of this data (Costello and Watts, 2016a). Similar results were obtained for other values of d. Table 3 gives the objective probabilities, the probability estimates from Experiments 1 and 2 in Zhao et al. (2009), and probability estimates produced by the simulation. As this table shows, estimates produced by the simulation were typically close to the average estimates from participants, and simulation estimates for the conjunction ratio $P(A \wedge B)/P(B)$ were significantly higher than simulation estimates for the conditional $P(A|B)$ at each probability level. Simulation estimates were biased away from objective probabilities in the same direction as seen in participant's estimates, and typically by the almost the same amounts.

At the very least, then, our model is consistent with these previous results on conditional probability estimation from experience, and explains the apparent contradiction between the results of Zhao et al. (2009) and Fisher and Wolfe (2014). In other work (Costello and Watts, 2016b) we have shown how this model can explain a series of results from Crupi et al. (2008) on the relationship between conditional probability values and the occurrence of the conjunction fallacy. In the next two sections we describe a series of much more stringent tests of our model, involving the identities in Tables 1 and 2.

5 Experiment 1

We now describe an experiment testing the predictions of our model; in particular, those concerning the identities shown in Tables 1 and 2. To test these predictions we gathered 62 participants' estimates for the 10 different constituent probability terms in the identities in Tables 1 and 2 (i.e. $P(A)$, $P(A \wedge B)$, $P(A|B)$ and so on) for five different pairs of events. We combined

Table 4: A, B weather event pairs used in the experiment.

pair	A, B pairs in Group 1	pair	A, B pairs in Group 2
1	cold, rainy	6	cloudy, rainy
2	cloudy, icy	7	cold, icy
3	cold, thundery	8	cloudy, thundery
4	cloudy, warm	9	sunny, warm
5	sunny, snowy	10	icy, snowy

each participant's individual estimates for each pair according to the given identities. Participants in Group 1 saw one set of five pairs of events and those in Group 2 saw a different set: we expected the predictions to hold for both groups.

5.1 Materials

We constructed two sets of pairs of weather events, each containing five pairs; participants in Group 1 gave estimates for one set of pairs and those in Group 2 gave estimates for the second set. The two sets are shown in Table 4; the pairs were selected so that each set contained events of high, medium and low probabilities, and with varying degrees of dependency between the elements of the pairs: some pairs had positive dependencies (it is more likely to be rainy if it is cloudy), some had negative dependencies (it is less likely to be snowy if it is sunny), and others were essentially independent.

5.2 Method

Participants were 62 undergraduate students at the School of Computer Science and Informatics, UCD, who volunteered to take part in exchange for partial course credit, and gave informed consent. Participants were asked to estimate the ten probabilities

$$P(A),\ P(B),\ P(A \wedge B),\ P(A \wedge \neg B),\ P(\neg A \wedge B),\ P(A \vee B),$$
$$P(A|B),\ P(B|A),\ P(A|\neg B),\ P(B|\neg A)$$

for each of the five pairs of weather events, giving 50 estimation questions for each participant. For single events, conjunctions, disjunctions and conjunctions with negations participants were asked questions of the form

- What is the probability that the weather will be W on a randomly-selected day in Ireland?

where the weather event W could be a single event such as 'cloudy', a conjunction such as 'cloudy and cold', a disjunction such as 'cloudy or cold' or a conjunction and negation such as 'cloudy and not cold' or 'cold and not cloudy'. For conditionals, participants were asked questions of the form

- If the weather in Ireland is W on a given randomly selected day, what is the probability that the weather will also be X on that same day?

where W and X were the two single component events of the conditional $P(X|W)$. Participants gave their estimates on a 100-point scale, with the 0 point labelled 'will never happen' and the 100 point labelled 'certain to happen'. Questions were presented in random order on a web browser. The task took around half an hour to complete. Participants' responses were divided by 100 prior to analysis.

5.3 Results

Two participants were excluded because they gave the same response for all questions, leaving 60 participants (31 in Group 1 and 29 in Group 2). As a check to ensure consistency across participants, we split participants in each group into two random groups and calculated the average probability estimate in each group for each one of the 50 presented probability terms. If participants were responding consistently we would expect there to be a high correlation between average estimates from one half of the participants and average estimates from the other half, in both groups. Correlations were high in both groups ($r = 0.96$, $p < 0.0001$ and $r = 0.97$, $p < 0.0001$), indicating consistent responses.

5.3.1 Deviations from probability theory

We expected significant deviations from probability theory in people's responses; this expectation was confirmed. Every participant committed the conjunction fallacy at least once and all but 4 committed the disjunction fallacy at least once. On average participants committed the conjunction fallacy for 56% of conjunctions and committed the disjunction fallacy for 49% of disjunctions: these fallacy rates represent significant deviations from the requirements of probability theory. Average values for the identities in Table 2 were positive for every A, B pair in both groups, again representing significant deviation from probability theory's requirement that these identities have a value of 0 (see Table 5). Average values for every one of these identities were significantly different from probability theory's value of 0 in one-sample t-tests across individual values in both groups ($t(154) > 9.97$, $p < 0.002$ for all identities in Group 1, $t(144) > 7.8$, $p < 0.002$ for all identities in Group 2, with Bonferroni correction for multiple comparisons), just as predicted by our model.

Recall that our model predicts that Identities 9 through 12 should all have the same average value, equal to d (the rate of random error for a given participant), and that Identities 13 through 16 should all have the same average value, equal to $d/2$. The average values in Table 5 support this prediction: values for Identities 9 through 12 were all close to their overall mean of 0.24 and values for Identities 13 through 16 were all around half that value (close to their overall mean of 0.12). Recall that our model also predicts that values for these identities for different events

Table 5: Average value (SD) for Identities 9 through 16 in Experiment 1, computed from participants' probability estimates in Groups 1 and 2. All are positive and significantly different from zero as predicted by our model ($p < 0.002$ in all cases in one-sample t-tests with Bonferroni correction for multiple comparisons). Note that the average values for identities 13 to 16 are approximately half the average values of identities 9 to 12, as predicted by our model.

	Group		
Identity	1	2	predicted value
9	0.24 (0.29)	0.17 (0.26)	d
10	0.25 (0.31)	0.25 (0.31)	d
11	0.25 (0.31)	0.28 (0.32)	d
12	0.24 (0.29)	0.20 (0.28)	d
13	0.14 (0.18)	0.10 (0.18)	$d/2$
14	0.13 (0.20)	0.10 (0.20)	$d/2$
15	0.12 (0.26)	0.12 (0.27)	$d/2$
16	0.13 (0.22)	0.12 (0.22)	$d/2$

should fall in the range $d^2 \ldots d(1-d)$. Taking d to have a value of 0.24 (the average value of Identities 9 through 12), we have a predicted range of $.06 \ldots .18$: all values for Identities 13 through 16 were well within that range.

We would expect this chance of random error, d, to vary across participants, but to be relatively constant within a given participant. To test this prediction, for each participant we calculated the average value of Identities 9 through 12 from that participant's estimates and measured the correlation between participants' values for pairs of identities. All correlations were positive (average pairwise correlation of $r = 0.57$); of the six pairs, five showed a significant correlation at the $p < 0.001$ level (with Bonferroni correction for multiple comparisons), while correlation for the remaining pair was not significant ($p = 0.17$). Similarly, for each participant we calculated the average value of Identities 13 through 16 from that participant's estimates and measured the correlation between participants' values for pairs of identities. Again, all correlations were positive (average pairwise correlation of $r = 0.71$); all pairs showed a significant correlation at the $p < 0.001$ level (with Bonferroni correction for multiple comparisons).

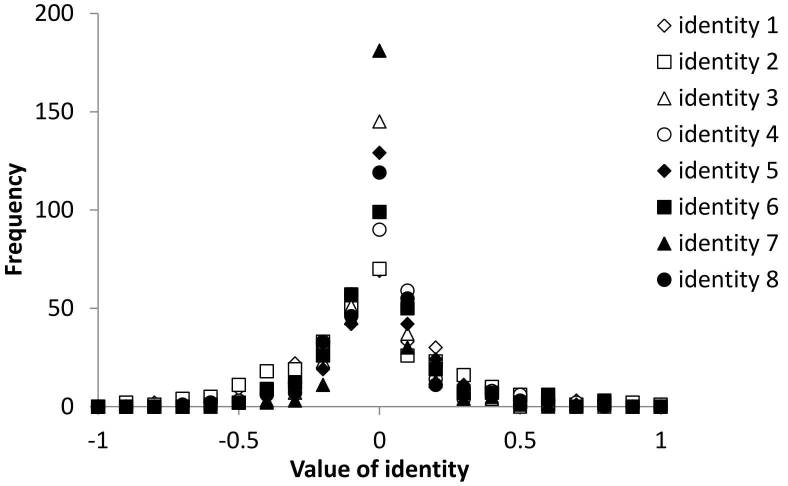

Figure 1: Frequency of occurrence of different values for Identities 1 through 8 in Experiment 1 across all A, B pairs in the experiment, grouped into 'bins' from $v - 0.05 \ldots v + 0.05$ for v from -1 to $+1$ in steps of 0.1. For example, since there were 60 participants in the experiment and each participant saw five pairs of events, the value of Identity 7 was calculated $5 \times 60 = 300$ times in total. Grouping these values into bins, we find that just over 180 of these calculations gave a value that fell in the $-0.05 \ldots + 0.05$ bin. Probability theory predicts these values will be symmetric around 0.

5.3.2 Agreement with probability theory

Our model predicts that average values for the identities in Table 1 will be close to 0, the value required by probability theory: this expectation was confirmed. For all of the identities in Table 1, participants' responses had an average value very close to 0 in both Groups 1 and 2. Averaging across all these identities gave a grand mean of $M = -0.006(95\%, SD = 0.22)$. Figure 1 graphs the frequency of occurrence of values for these identities. It is clear from the graph that values for these identities are approximately symmetrically distributed around 0, the value predicted by our model.

Table 6: Average value (SD) for Identities 1 through 8 in Groups 1 and 2, Experiment 1, computed from participants' probability estimates. All are close to zero, as predicted by our model (of the 80 different averages, 74, or 92.5%, fell in the range $-0.1 \ldots + 0.1$). Out of 80 separate one-sample t-tests on the individual values making up these averages, only one was marginally significant ($p = 0.04$ with Bonferroni correction for multiple comparisons).

Group 1

	pair					
Identity	1	2	3	4	5	overall
1	-0.07 (0.21)	0.08 (0.4)	-0.07 (0.32)	0.03 (0.36)	0.03 (0.20)	0.00 (0.31)
2	-0.12 (0.27)	0.05 (0.25)	0.09 (0.28)	0.02 (0.27)	-0.08 (0.19)	-0.01 (0.26)
3	0.05 (0.18)	-0.02 (0.09)	-0.02 (0.10)	-0.08 (0.13)	0.00 (0.06)	-0.01 (0.12)
4	0.01 (0.20)	-0.07 (0.14)	-0.02 (0.25)	-0.11 (0.20)	0.06 (0.19)	-0.02 (0.20)
5	-0.08 (0.22)	0.00 (0.14)	0.05 (0.20)	0.05 (0.18)	0.01 (0.19)	0.01 (0.19)
6	-0.04 (0.19)	-0.05 (0.17)	0.00 (0.26)	-0.03 (0.21)	0.06 (0.20)	-0.01 (0.21)
7	-0.03 (0.18)	-0.02 (0.09)	0.04 (0.16)	-0.03 (0.14)	0.01 (0.18)	-0.01 (0.16)
8	0.04 (0.13)	-0.04 (0.15)	-0.06 (0.02)	-0.08 (0.12)	0.05 (0.16)	-0.02 (0.16)

Group 2

	pair					
Identity	6	7	8	9	10	overall
1	-0.02 (0.22)	-0.07 (0.23)	0.04 (0.29)	-0.04 (0.28)	-0.05 (0.25)	-0.03 (0.26)
2	-0.21 (0.34)	-0.03 (0.27)	0.02 (0.27)	-0.12 (0.23)	-0.08 (0.35)	-0.08 (0.30)†
3	0.07 (0.15)	-0.05 (0.15)	-0.07 (0.15)	0.05 (0.13)	0.01 (0.15)	0.00 (0.16)
4	-0.01 (0.14)	-0.04 (0.18)	-0.08 (0.18)	0.16 (0.22)	0.06 (0.19)	0.02 (0.20)
5	-0.10 (0.13)*	0.08 (0.20)	0.07 (0.19)	0.01 (0.14)	0.05 (0.23)	0.02 (0.19)
6	-0.07 (0.12)	0.01 (0.18)	0.00 (0.25)	0.11 (0.21)	0.05 (0.16)	0.02 (0.20)
7	-0.03 (0.11)	0.03 (0.10)	-0.01 (0.06)	0.06 (0.12)	0.07 (0.13)	0.02 (0.11)
8	0.02 (0.09)	-0.07 (0.21)	-0.07 (0.17)	0.10 (0.20)	0.00 (0.20)	-0.00 (0.19)

* $p = 0.04$, with Bonferroni correction for multiple comparisons.

† Evidence in favour of the alternative in a JZS Bayes Factor test (JZS Bayes Factor $= 18.6$). Overall values for all other identities gave evidence in favour of the null hypothesis (values of 0).

This pattern also held for each individual event pair A, B. Tables 6 shows the average values for these identities obtained for each event pair in Group 1 and Group 2. There are 80 different averages across these two tables (10 event pairs by 8 identities); of these 74 (92.5%) fell in the range $-0.1 \ldots + 0.1$, and 48 (60%) fell in the range $-0.05 \ldots + 0.05$. We analysed the distribution of these values for individual event pairs by carrying out 80 separate one-sample t-tests. Of these 80 tests, only one was marginally significant ($p = 0.04$ with Bonferroni correction for multiple comparisons). JZS Bayes Factor tests on overall values for identities (across all pairs) gave evidence in favour of the null hypothesis (that the value for the identity is 0) in all but one identity (and the overall value for that identity was still relatively close to 0). These results suggest that values for all these identities are distributed around 0, just as predicted by our model.

Finally, we carried out a more detailed analysis of the distribution of values for these identities by producing, for each identity 1 through 8, a scatterplot relating the positive and negative terms of the given identity calculated from a single participant's estimates for a single event pair A, B. For example, for identity 1 each point in this scatterplot represents on the x-axis the sum of a single participant's estimate for $P(A \wedge B)$ and for $P(A \vee B)$ for a single pair for events A, B, and on the y-axis the sum of the same participant's estimate for $P(A)$ and for $P(B)$ for the same events.

These eight scatterplots are shown in Figure 2. Since there were 60 participants in total, and each participant gave estimates for five A, B pairs, there are 300 points in each of these scatterplots. Our model predicts these points will be symmetrically distributed around the line of equality (dashed). Lines of best fit to the data (solid) were obtained via Deming regression, which accounts for error in observations on both the x- and y-axes. These lines of best fit represent maximum likelihood estimators for the relationship between the x and y variables given random error in both variables: in other words, these lines of best fit represent the most

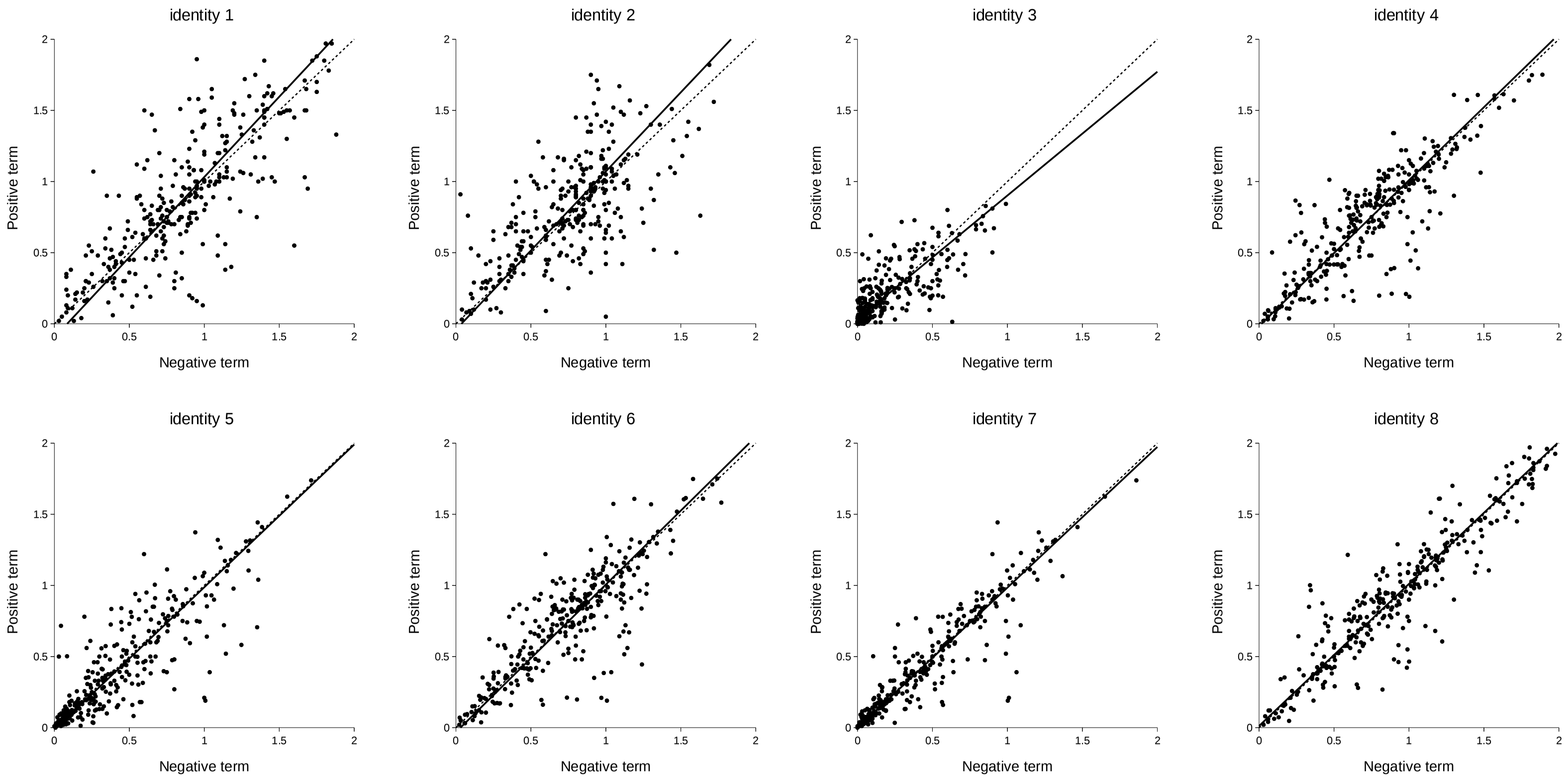

Figure 2. Scatterplots of values for the positive and negative terms of Identities 1 through 8 in Experiment 1. Each point represents the value of the positive and negative terms of the given identity calculated from a single participant's estimates for a single event pair A, B. For example, for Identity 1 (the addition law) each point represents on the x-axis the sum of a single participant's estimate for $P(A \wedge B)$ and for $P(A \vee B)$ for a single pair for events A, B, and on the y-axis the sum of the same participant's estimate for $P(A)$ and for $P(B)$ for the same events. Our model predicts these points will be symmetrically distributed around the line of equality (dashed), which represents the value required by probability theory. Lines of best fit to the data (solid) were obtained via Deming regression, which accounts for error in observations on both the x- and y-axes.

likely relationship between the x and y values, given the observed set of data points and the assumption of random error.

It is clear from these eight scatterplots that participants' individual values for these identities were distributed along and approximately symmetrically around the line of equality, as predicted by the model: G_1 sample skewness for values for each identity in this figure all fell in the range ± 0.15 indicating symmetric distributions (Bulmer, 2012). These scatterplots thus show in detail the agreement between people's probability estimates and the requirements of probability theory for these identities. In particular, the fact that the lines of best fit in these scatterplots are very close to the line of equality indicates that, in individual participants' responses for individual event pairs, the most likely relationship between positive and negative terms for these identities (between $P_E(A) + P_E(B)$ and $P_E(A \wedge B) + P_E(A \vee B)$ for identity 1, for example) is equality, just as required by probability theory. People's probability estimates, when combined in these identities, are surprisingly rational.

5.4 Discussion

What is the scope of this model of people's probabilistic reasoning? Since the model assumes that the probability $P(A|B)$ is estimated by retrieving a sample of specific instances of B from memory and counting the number of A's, it may seem at first glance that the model is only able to give probability estimates for events that have already been seen. This position depends on a conception of memory as being nothing but a store of recorded events. We can, however, take an alternative conception of memory as a constructive process that can generate representations of events even if those specific events have not previously been seen. Support for this view comes from evidence that remembering past events and imagining future events are very similar cognitive processes (see e.g. Schacter, 2012).

If we take this 'constructive' or 'simulation' view of memory then our model can apply to probability estimates for all forms of event, whether previously seen or completely novel. In this view, to estimate $P(A|B)$ a sample of specific instances of B are generated by constructive memory, and then the probability of A is estimated by counting the number of those instances that are A's. Random noise in the counting process causes the observed patterns of bias and agreement with probability theory in these estimates, as described by the model.

Under this view we would expect to see similar agreement with probability theory for identities such as those in Table 1 in situations where we ask people to estimate probabilities for events which they have repeatedly experienced in the past (such as the weather events in our experiment), for past events which they have not directly experienced, for events which are to some degree imaginary, and even for possible unique events in the future. We would also expect the degree of random error d to vary across these different types of event; for example, we would expect a higher degree of random error for unique future events (which necessarily have a high degree of associated uncertainty) than for repeatedly experienced past events.

Experimental results support these predictions, at least for some of the identities in the table. For example, in an experiment asking participants to estimate the probability of people over the age of 60 having certain diseases (such as Alzheimer's, diabetes and so on), having conjunctions of those diseases (Alzheimer's and diabetes) and having disjunctions of those diseases (Alzheimer's or diabetes), Costello and Mathison (2014) found that people's probability estimates agreed very closely with identity 1, the addition law. In an experiment asking participants to estimate the probability of a range of future events (such as a future increase in cigarette taxes, a future decline in smoking rates, and so on) and of various conjunctions and disjunctions of those events, we found that people's probability estimates for these future events agreed very closely with Identities 1 and 2 and with a number of other such identities derived from our model (Costello and Watts, 2016c). Further the estimated values of d obtained for these events were noticably higher than those obtained for the weather events in the experiment described here. Finally, in an experiment where participants were given personality descriptions for a range of imaginary people and then asked to assess the probability of various direct, conjunctive, disjunctive and conditional statements being true for those people, Fisher and Wolfe (2014) found that people's probability estimates for these imaginary events agreed very closely with identities 1 (the addition law) and 3 (the addition form of Bayes' rule). Together, these results suggest that our model could potentially apply to probability estimation for events in general, and is not limited solely to events that have previously been seen. In the next section we describe an experiment which investigates this possibility.

6 Experiment 2

The results of Experiment 1 confirmed our model's predictions concerning the identities in Tables 1 and 2. These results, however, only apply to events for which participants have repeated and everyday experience (weather events). We now describe an experiment testing whether the predictions of our model will extend to novel, unique events for which participants have no everyday experience. Since these events are more complex than the simple and familiar weather events used in Experiment 1, we expect the noise rate d to be higher for these events. However, we still expect the model's predictions to hold.

To test these predictions we gathered 70 participants' estimates for the various constituent probability terms in the identities in Tables 1 and 2 (i.e. $P(A)$, $P(A \wedge B)$, $P(A|B)$ and so on) for three different pairs of novel events that could potentially occur in the future. We combined each participant's individual estimates for each pair according to the given identities. Participants in Group 1 saw one set of three pairs of events and those in Group 2 saw a different set: we expected the predictions to hold for both groups.

Table 7: Sets of A,$\neg A$ and B,$\neg B$ events used in experiment 2. For each event-set these basic events were used to construct following complex events: $A \wedge B$, $A \wedge \neg B$, $\neg A \wedge B$, $\neg A \wedge \neg B$, $A \vee B$, $A|B$, $B|A$, $A|\neg B$, $B|\neg A$. Participants in Group 1 were asked about all simple and complex events from sets $1, 2$ and 3: participants in Group 2 were asked about all simple and complex events from sets $4, 5$ and 6.

event-set	A event	$\neg A$ event	B event	$\neg B$ event
1	Britain has left the European Union	Britain has NOT left the European Union	Greece has left the European Union	Greece has NOT left the European Union
2	The number of cars on Irish roads has increased	The number of cars on Irish roads has NOT increased	The price of petrol has increased significantly	The price of petrol has NOT increased significantly
3	Climate change has a large impact on Ireland's weather	Climate change does NOT have a large impact on Ireland's weather	World greenhouse gas emissions have been reduced	World greenhouse gas emissions have NOT been reduced
4	The US is at war in the Middle East	The US is NOT at war in the Middle East	There has been another major terrorist attack in the US	There has NOT been another major terrorist attack in the US
5	Europe has grown noticably poorer	Europe has NOT grown noticably poorer	Unemployment in Europe is above 20%	Unemployment in Europe is NOT above 20%
6	Hurricanes and typhoons have become more frequent	Hurricanes and typhoons have NOT become more frequent	The average world temperature has increased	The average world temperature has NOT increased

6.1 Materials

We constructed two sets of pairs of novel future events, each containing three pairs; participants in Group 1 gave estimates for one set of pairs and those in Group 2 gave estimates for the second set. The two sets are shown in Table 7; the pairs were selected so that each set contained events of high, medium and low probabilities, and all with some degree of conditional or causal relationship between the elements of the pairs, asking, for example, about reductions in greenhouse gas emissions and about the impact of climate chance. We used such causally linked pairs for two reasons: first, to ensure that the questions about conditional probabilities were natural and easy to understand for participants (when two events are causally linked, questions about the conditional probability of one event given the other arise naturally); and second, to address any concern that our results from Experiment 1 could have occurred because the weather events used in that experiment may not have had strong conditional relationships.

6.2 Method

Participants were 70 undergraduate students at the School of Computer Science and Informatics, UCD, who volunteered to take part in exchange for partial course credit, and gave informed consent. Participants were asked to estimate the probabilities

$$P(A),\ P(B),\ P(A \wedge B),\ P(A \wedge \neg B),\ P(\neg A \wedge B),\ P(\neg A \wedge \neg B),\ P(A \vee B),$$
$$P(A|B),\ P(B|A),\ P(A|\neg B),\ P(B|\neg A)$$

for each of the three pairs of events, giving 33 estimation questions for each participant. For single events, conjunctions, disjunctions and conjunctions with negations participants were asked questions of the form

- What is the probability that X by the year 2025?

where the event X could be a single event such as ‘Britain has left the European Union’, a conjunction such as ‘Britain has left the European Union **and** Greece has left the European Union’, or some other such combination. For conditionals, participants were asked questions of the form

- Imagine that Y by the year 2025. If that happens, then what is the probability that X by the year 2025?

where Y and X were the two single component events of the conditional $P(X|Y)$. Participants gave their estimates on a 100-point scale, with the 0 point labelled ‘will never happen’ and the 100 point labelled ‘certain to happen’. Questions were presented in random order on a web browser. The task took around half an hour to complete. Participants’ responses these were divided by 100 prior to analysis.

6.3 Results

Four participants were excluded because due to a computer error they did not see all questions, leaving 66 participants (32 in Group 1 and 34 in Group 2). As a consistency check we split participants in each group into two random groups and calculated the average probability estimate in each group for each one of the 33 presented probability terms. If participants were responding consistently we would expect there to be a reliable correlations between these split-half averages. Both groups showed a high correlation between split-half averages ($r = 0.90$, $p < 0.0001$ and $r = 0.92$, $p < 0.0001$), indicating consistent responses.

6.3.1 Deviations from probability theory

We expected significant deviations from probability theory in people’s responses; again, this expectation was confirmed. 57 out of the 66 participants committed the conjunction fallacy at

Table 8: Average value (SD) for Identities 9 through 16, computed from participants' probability estimates for all event sets in Experiment 2. All are positive and significantly different from zero as predicted by our model ($p < 0.002$ in all cases in one-sample t-tests with Bonferroni correction for multiple comparisons). Note that, as in Experiment 1, the average values for identities 13 to 16 are approximately half the average values of identities 9 to 12, as predicted by our model.

Identity	average value	predicted value
9	0.37 (0.29)	d
10	0.31 (0.29)	d
11	0.38 (0.29)	d
12	0.43 (0.30)	d
13	0.20 (0.21)	$d/2$
14	0.17 (0.20)	$d/2$
15	0.16 (0.21)	$d/2$
16	0.14 (0.24)	$d/2$

least once and 48 committed the disjunction fallacy at least once. On average participants committed the conjunction fallacy for 56% of conjunctions and committed the disjunction fallacy for 34% of disjunctions. As Table 8 shows, average values for the identities in Table 2 were positive and significantly different from 0 for every identity ($t(95) > 4.94$, $p < 0.002$ for all identities in Group 1, $t(101) > 6.68$, $p < 0.002$ for all identities in Group 2, in one-sample t-tests with Bonferroni correction for multiple comparisons), just as predicted by our model.

Recall that our model predicts that Identities 9 through 12 should all have the same average value, equal to d (the rate of random error for a given participant), and that Identities 13 through 16 should all have the same average value, equal to $d/2$. The average values in Table 8 support this prediction: values for Identities 9 through 12 were all close to their overall mean of 0.37 and values for Identities 13 through 16 were all around half that value. Recall that our model also predicts that values for these identities for different events should fall in the range $d^2 \ldots d(1-d)$. Taking d to have a value of 0.37 (the average value of Identities 9 through 12), we have a predicted range of $.14 \ldots .23$: all values for Identities 13 through 16 were within that range.

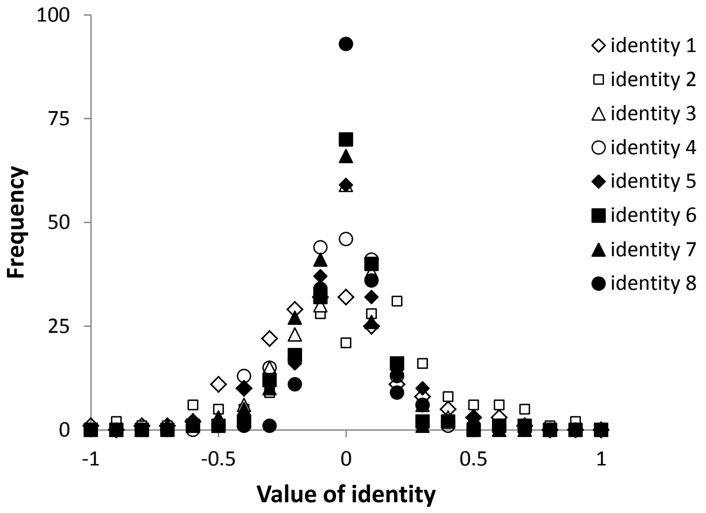

Figure 3: Frequency of occurrence of different values for Identities 1 through 8 in Experiment 2 across all A, B pairs in the experiment, grouped into 'bins' from $v - 0.05 \ldots v + 0.05$ for v from -1 to $+1$ in steps of 0.1.

Note that the average values for d in this experiment was higher than the average value seen in Experiment 1, reflecting higher noise rates for these 'future event' materials than for the 'weather event' materials used in that experiment. This difference is just as we would expect given that these future events were significantly more complex than the weather events used in Experiment 1.

6.3.2 Agreement with probability theory

Again, our model predicts that average values for the identities in Table 1 will be close to 0, the value required by probability theory: this expectation was confirmed. Averaging across all these identities gave a grand mean of $M = -0.02(95\%, SD = 0.22)$. Figure 3 graphs the frequency of occurrence of values for these identities. As in Experiment 1, values for these identities are symmetrically distributed around 0, just as predicted by our model.

This pattern also held for each individual event pair A, B. Table 9 show the average values for these identities obtained for each event pair. There are 48 different averages in this table; of these 40 (83.3%) fell in the range $-0.1 \ldots + 0.1$, and 29 (60%) fell in the range $-0.05 \ldots + 0.05$. We analysed the distribution of these values for individual event pairs by carrying out 48 separate one-sample t-tests. Of these 48 tests, just 3 t-tests were significant at the $p < 0.05$ level (with Bonferroni correction for multiple comparisons). JZS Bayes Factor tests on overall values for identities (across all pairs) gave evidence in favour of the null hypothesis (that the value for the identity is 0) in all but 3 identities (and the overall values for those identities were still close to 0). These results suggest that values for these identities are distributed around 0, just as predicted by our model. Of course, the fact that some values for these identities were different from 0 means that there may be some other factor in play that is not accounted for in our model. However, even values that had statistically-significantly differences from 0 were nevertheless still close to 0. This suggests that, even if there is some other such factor, the influence it has is small.

As before, we carried out a more detailed analysis of the distribution of values for these identities by producing, for each identity 1 through 8, a scatterplot relating the positive and negative terms of the given identity calculated from a single participant's estimates for a single event pair A, B (see Figure 4). Since there were 66 participants in total, and each participant gave estimates for three A, B pairs, there are 198 points in each scatterplot. Our model predicts these points will be symmetrically distributed around the line of equality (dashed), which represents the value required by probability theory. As before, lines of best fit to the data (solid) were obtained via Deming regression: these lines of best fit represent the most likely relationship between the x and y values, given the observed set of data points and the assumption of random error. As in Experiment 1, it is clear from these eight scatterplots that participants' individual

Table 9: Average value (SD) for Identities 1 through 8, computed from participants' probability estimates for events in event sets 1 to 6 in Experiment 2. All are close to zero, as predicted by our model (of the 48 different averages, 40, or 83.3%, fell in the range $-0.1 \ldots + 0.1$). Out of 48 separate one-sample t-tests on the individual values making up these averages, 3 were significantly different from 0 at the $p < 0.05$ level (with Bonferroni correction for multiple comparisons).

	event set						
Identity	1	2	3	4	5	6	overall
1	-0.04 (0.26)	-0.05 (0.25)	-0.07 (0.30)	-0.03 (0.26)	-0.19 (0.26)*	-0.03 (0.34)	-0.07 (0.28)†
2	0.00 (0.23)	0.19 (0.30)	0.02 (0.35)	-0.12 (0.36)	0.14 (0.30)	0.09 (0.30)	0.05 (0.32)
3	-0.04 (0.11)	-0.12 (0.19)	-0.06 (0.18)	0.01 (0.16)	-0.07 (0.18)	0.08 (0.18)	-0.03 (0.18)
4	0.03 (0.16)	-0.16 (0.17)*	-0.08 (0.13)	0.00 (0.14)	-0.03 (0.19)	-0.02 (0.19)	-0.04 (0.18)†
5	-0.02 (0.16)	-0.04 (0.21)	-0.03 (0.25)	-0.03 (0.18)	0.10 (0.14)	-0.12 (0.20)	-0.02 (0.20)
6	0.07 (0.14)	-0.04 (0.18)	-0.02 (0.21)	-0.02 (0.13)	0.05 (0.10)	-0.10 (0.14)	-0.01 (0.16)
7	-0.05 (0.14)	-0.16 (0.15)*	-0.09 (0.17)	-0.01 (0.14)	0.03 (0.16)	-0.05 (0.19)	-0.06 (0.17)†
8	0.08 (0.17)	0.00 (0.08)	0.01 (0.15)	0.01 (0.11)	-0.05 (0.14)	0.02 (0.16)	0.01 (0.14)

* $p < 0.05$, with Bonferroni correction for multiple comparisons.

† Evidence in favour of the alternative in a JZS Bayes Factor test (JZS Bayes Factor > 16.2). Overall values for all other identities gave evidence in favour of the null hypothesis.

values for these identities were distributed along and symmetrically around the line of equality, as predicted by the model: G_1 sample skewness for values of each identity all fell in the range ± 0.09, indicating symmetric distributions (Bulmer, 2012). Again, the fact that the lines of best fit in these scatterplots are very close to the line of equality indicates that, in individual participants' responses for individual event pairs, the most likely relationship between positive and negative terms for these identities is equality, just as required by standard probability theory and just as predicted by our model.

7 General Discussion

We can summarise the main point of our work as follows: when deviations due to noise are cancelled out in people's probability judgments (as in Identities 1 through 8), those judgements are, on average, just as required by probability theory with no systematic bias. This average agreement with probability holds for a range of different individual events (as shown in Tables 6 and 9), for familiar repeated events (as shown in Experiment 1) and for unfamiliar unique events (as shown in Experiment 2), and holds, on average, at the level of individual people's responses (as shown in the scatterplots). This close agreement with probability theory cannot be dismissed by suggesting that our participants happened to be particularly good at probability estimation, because this agreement occurs alongside significant rates of conjunction and disjunction fallacy occurrence in the same participants' responses for the same events, and alongside significant bias away from the requirements of probability theory for identities which do not cancel out the effects of random noise (Identities 9 through 16). For these identities the average degree of bias follows the predictions of our model (an approximately constant degree of bias d for Identities 9 through 12, and an approximately constant degree of bias $d/2$ for Identities 13 through 16). Taken together, the most natural explanation for these results seems to be that people estimate probabilities using a mechanism that is fundamentally rational (in line with

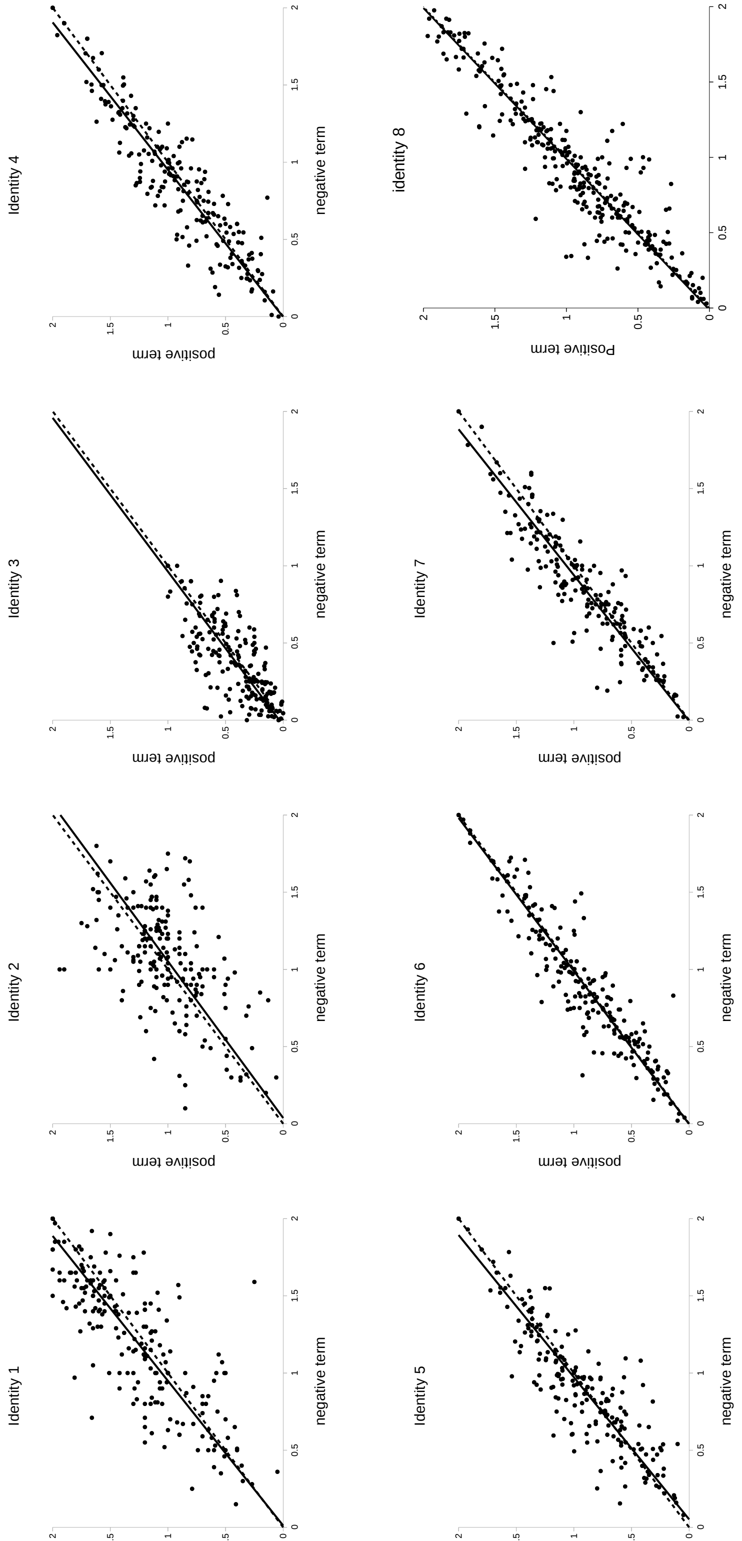

Figure 4. Scatterplots of values for the positive and negative terms of Identities 1 through 8 in Experiment 2. Each point represents the value of the positive and negative terms of the given identity calculated from a single participant's estimates for a single event pair A, B. For example, for Identity 1 (the addition law) each point represents on the x-axis the sum of a single participant's estimate for $P(A \wedge B)$ and for $P(A \vee B)$ for a single pair for events A, B, and on the y-axis the sum of the same participant's estimate for $P(A)$ and for $P(B)$ for the same events. Our model predicts these points will be symmetrically distributed around the line of equality (dashed). Lines of best fit to the data (solid) were obtained via Deming regression, which accounts for error in observations on both the x- and y-axes.

frequentist probability theory), but is subject to the biasing effects of random noise.

7.1 Rationality and noise

Our theoretical proposal here is that human probabilistic reasoning is based on a fundamentally rational process that is subject to random noise. It is important to stress that we are not suggesting that people are consciously aware of the equations of probability theory when estimating probabilities. That is clearly not the case, given the high rates of conjunction fallacy occurrence in people's judgments for some events. Instead we propose that people's probability judgments are derived from a 'black box' that estimates the probability of an event by retrieving (some analogue of) a count of instances of that event from memory. Such a mechanism is necessarily subject to the requirements of set theory and therefore embodies the rules of probability theory.

We expect this probability module to be unconscious, automatic, rapid, parallel, relatively undemanding of cognitive capacity and evolutionarily 'old'. Support for this view comes from that fact that people make probability judgments rapidly and easily and typically do not have access to the reasons behind their estimations, from extensive evidence that event frequencies are stored in memory by an automatic and unconscious encoding process (Hasher and Zacks, 1984) and from evidence suggesting that infants have surprisingly sophisticated representations of probability (Cesana-Arlotti et al., 2012). Other support comes from results showing that animals effectively judge probabilities (for instance, the probability of obtaining food from a given source) and that their judged probabilities are typically close to optimal (Kheifets and Gallistel, 2012).

It is equally important to stress that we are not suggesting that people's probability estimates are themselves rational. Again, this is clearly not the case: there is very extensive evidence demonstrating that people's probability estimates are systematically biased away from the requirements of probability theory. We argue that these biases are a consequence of the influence of random noise on the probability estimates generated by an underlying rational process. While this noise is random, it has systematic, directional effects: for example, our noisy model's expected averages for probability estimates are systematically biased away from the 'true' probability values, in a way that seems to match the biases seen in people's estimates.

It is useful to expand on the distinction between the rationality of a process (for probability estimation) and the rationality of the outputs (the probability estimates) produced by that process. Some might argue that is wrong to classify a reasoning process as 'rational' when the outputs it produces are systematically biased away from the objectively correct, rational requirements. We feel that this argument holds only in a perfectly noise-free situation: if there were no noise in reasoning, then we would indeed expect a rational process to produce outputs that exactly match the objectively correct rational requirements in all cases, and we would classify a process as irrational if its outputs deviated from those rational requirements in any way. When we consider the problem of reasoning in a noisy system, however, the position is different. Here, no reasoning process can meet the strict criteria for rationality: no process can produce outputs that match objectively correct rational requirements in all cases (because every process is subject to random error). Given that the presence of noise puts limits on the extent to which any process can approach 'perfect' rationality (the more noise is present, the more every process will be subject to error), we need a more subtle criterion for the rationality of a reasoning process. A natural criterion is one which classifies a process as rational if the outputs from that process come to match the objectively correct rational requirements more and more closely as the degree of noise in the reasoning process falls, with a perfect match when noise falls to zero. Our model exactly satisfies this requirement, because it reduces to standard probability theory when the noise term d is zero.

7.2 Implications for other models of probabilistic reasoning

Our results have implications for current approaches to the psychology of people's probabilistic reasoning. In particular, our results are problematic for the view that people estimate probabilities via heuristics that do not 'follow the calculus of chance or the statistical theory of prediction' (Kahneman and Tversky, 1973, p. 237). It seems clear to us that such heuristic accounts are motivated by the assumption that the observed biases and errors seen in people's probability judgments cannot be explained by probability theory. This motivation arises because probability theory is the normative model against which these biases and errors are assessed. If researchers had not taken those biases and errors as evidence that people don't reason using probability theory, they would have had no reason to propose those alternative accounts. However, our model suggests that these biases do not, in fact, count as evidence that people don't reason using probability theory. Those alternative models thus lose their fundamental motivation: there is no reason for moving from probability theory to those alternative accounts in an attempt to explain human probabilistic reasoning. There is, in contrast, an underlying motivation for the probability theory plus noise model: the probability of events in the world necessarily follow the rules of probability theory, and our reasoning processes are necessarily subject to noise.

Our results also have implications for the currently popular proposal that people follow normative models of reasoning based on Bayesian inference, a process for drawing conclusions given observed data in a way that follows probability theory (Tenenbaum et al., 2011, Chater et al., 2006, Oaksford and Chater, 2007, Griffiths and Tenenbaum, 2006). Bayesian inference applies to conditional probabilities such as the probability of some conclusion H given some evidence E: $P(H|E)$. In Bayesian models these conditional probabilities are computed according to Bayes' rule

$$P(H|E) \quad = \quad \frac{P(H)P(E|H)}{P(E)}$$

and so the value of the conditional probability $P(H|E)$ depends on the value of the 'prior' $P(H)$ (the probability of H being true independent of the evidence E) and on the value of the 'likelihood function' $P(E|H)$ (the probability of seeing evidence E given that the hypothesis H is true).

At first glance it may seem that our experimental results support the Bayesian proposal: our results, after all, show that people's conditional probability estimates, on average, agree with the addition form of Bayes' rule. More fundamentally, however, our approach goes against the idea of Bayesian inference as the source of conditional probability estimates. This is because our model takes a 'frequentist' approach to probabilistic reasoning, where conditional probabilities are estimated directly by counting event occurrence in recalled items. Bayesian accounts, by contrast, see conditional probabilities as being produced indirectly, via computation from other probability estimates (the priors). Apart from the various problems with this assumption that other prior probabilities underlie probability estimation (see, e.g. Bowers and Davis, 2012, Marcus and Davis, 2013, Eberhardt and Danks, 2011, Jones and Love, 2011, Endress, 2013), it seems to us that this Bayesian approach would make predictions that diverge from those of our model, especially in the case of the various 'biased' identities given in Table 2. For these identities the Bayesian approach would seem to predict agreement with probability theory, rather than deviation from probability theory as predicted by our model (and as seen in our experiments).

We finally consider other models where people follow probability theory in their probabilistic reasoning, and where reasoning is based on the recall of items from memory subject to the biasing effects of random noise; models such as Minerva-DM (Dougherty et al., 1999) and Hilbert's 'noisy channel' model (Hilbert, 2012). Those models are both quite complex, containing a number of different mechanisms and parameters and identifying a number of different points at which noise can affect probabilistic estimation. Because of this complexity, it is difficult to derive clear and testable predictions from these models. This is the main advantage of our account: its simplicity allows us to derive clear, specific and verifiable predictions about the impact of random variation on human probabilistic reasoning. While our results give general support for the approach taken in these models, they neither support nor contradict the various detailed mechanisms for noise proposed in those models. An important aim for future work is to examine ways of investigating and distinguishing between these different mechanisms for noise.

8 Conclusions

Our results support the classical view of probability theory as ‘common sense reduced to calculus’ (Laplace, 1820), by showing that people’s probability estimates agree with fundamental requirements of probability theory such as the addition form of Bayes’ rule. The fact that this agreement occurs alongside significant deviations from probability for other expressions supports a model where people follow probability theory but are subject to the biasing effects of random error. These results go against the popular idea that people estimate probabilities via heuristic shortcuts that do not follow probability theory.

Our results have broader implications for research on patterns of bias in aspects of people’s decision-making. A common pattern in such research is to identify a systematic bias in people’s responses, and to then take that bias as evidence that people are reasoning via some heuristic shortcut rather than the correct reasoning process. Our results, however, show that this inference from observed bias to inferred heuristic can be premature: random noise in reasoning can cause systematic biases in people’s responses even when people are using normatively correct reasoning processes. To demonstrate conclusively that people are using heuristics, researchers must show that observed biases cannot be explained as the result of systematic effects caused by random noise.

References

Ariely, D. (2009). *Predictably irrational: the hidden forces that shape our decisions*. Harper-Collins.

Bondt, W. and Thaler, R. (2012). Does the stock market overreact? *The Journal of Finance*, 40(3):793–805.

Bowers, J. S. and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389.

Bulmer, M. G. (2012). *Principles of statistics*. Courier Corporation.

Camerer, C., Loewenstein, G., and Rabin, M. (2003). *Advances in Behavioral Economics*. Princeton University Press.

Cesana-Arlotti, N., Téglás, E., Bonatti, L. L., et al. (2012). The probable and the possible at 12 months: intuitive reasoning about the uncertain future. *Advances in Child Development and Behavior*, 43:1–25.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Costello, F. and Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3):463–480.

Costello, F. and Watts, P. (2016a). Explaining high conjunction fallacy rates: the probability theory plus noise account. *Journal of Behavioral Decision Making*. In press, available at http://dx.doi.org/10.1002/bdm.1936.

Costello, F. and Watts, P. (2016b). Probability theory plus noise: replies to Crupi and Tentori (2015) and to Nilsson, Juslin and Winman (2015). *Psychological Review*, 123(1):112–123.

Costello, F. and Watts, P. (2016c). Surprising rationality in people's probability estimation: Assessing two competing models of probability judgment. Submitted for publication.

Costello, F. J. and Mathison, T. (2014). On fallacies and normative reasoning: when people's judgements follow probability theory. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, pages 361–366.

Crupi, V., Fitelson, B., and Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, 14(2):182–199.

Dougherty, M. R. P., Gettys, C. F., and Ogden, E. E. (1999). Minerva-DM: A memory processes model for judgments of likelihood. *Psychological Review*, 106(1):180–209.

Eberhardt, F. and Danks, D. (2011). Confirmation in the cognitive sciences: the problematic case of bayesian models. *Minds and Machines*, 21(3):389–410.

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, 127(2):159 – 176.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3):519–527.

Eva, K. W. and Norman, G. R. (2005). Heuristics and biases: biased perspective on clinical reasoning. *Medical Education*, 39(9):870–872.

Fiedler, K., Brinkmann, B., Betsch, T., and Wild, B. (2000). A sampling approach to biases in conditional probability judgments: beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, 129(3):399.

Fisher, C. R. and Wolfe, C. R. (2014). Are people naïve probability theorists? A further examination of the probability theory + variation model. *Journal of Behavioral Decision Making*, 27(5):433–443.

Fisk, J. E. and Pidgeon, N. (1996). Component probabilities and the conjunction fallacy: Resolving signed summation and the low component model in a contingent approach. *Acta Psychologica*, 94(1):1–20.

Fox, C. R. and Levav, J. (2004). Partition-edit-count: naive extensional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, 133(4):626.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62:451–482.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102(4):684.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773.

Hasher, L. and Zacks, R. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *The American Psychologist*, 39(12):1372–1388.

Hertwig, R. and Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12):517–523.

Hicks, E. and Kluemper, G. (2011). Heuristic reasoning and cognitive biases: Are they hindrances to judgments and decision making in orthodontics? *American Journal of Orthodontics and Dentofacial Orthopedics*, 139(3):297–304.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, 138(2):211–237.

Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(04):169–188.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4):237.

Kahneman, D. and Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kheifets, A. and Gallistel, C. R. (2012). Mice take calculated risks. *Proceedings of the National Academy of Sciences*, in press.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(01):1–17.

Laplace, P. S. d. (1820). *Théorie analytique des probabilités*. Courcier.

Marcus, G. F. and Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24(12):2351–2360.

Oaksford, M. and Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford University Press.

Schacter, D. L. (2012). Adaptive constructive processes and the future of memory. *American Psychologist*, 67(8):603.

Shafir, E. and Leboeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, 53(1):491–517.

Sunstein, C. (2000). *Behavioral Law and Economics*. Cambridge University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273.

Williams, B. (2010). Heuristics and biases in military decision making. Technical report, DTIC Document.

Zhao, J., Shah, A., and Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, 113(1):26–36.