*Journal of Behavioral Decision Making*, *J. Behav. Dec. Making*, **30**: 304–321 (2017)
Published online 9 February 2016 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1936

# Explaining High Conjunction Fallacy Rates: The Probability Theory Plus Noise Account

FINTAN COSTELLO[1]* and PAUL WATTS[2]

[1] *School of Computer Science and Informatics, University College Dublin, Belfield, Ireland*

[2] *Department of Mathematical Physics, National University of Ireland Maynooth, Maynooth, Ireland*

## ABSTRACT

The conjunction fallacy occurs when people judge the conjunctive probability $P(A \wedge B)$ to be greater than a constituent probability *P(A)*, contrary to the norms of probability theory. This fallacy is a reliable, consistent and systematic part of people's probability judgements, attested in many studies over at least 40 years. For some events, these fallacies occur very frequently in people's judgements (at rates of 80% or more), while for other events, the fallacies are very rare (occurring at rates of 10% or less). This wide range of fallacy rates presents a challenge for current theories of the conjunction fallacy. We show how this wide range of observed fallacy rates can be explained by a simple model where people reason according to probability theory but are subject to random noise in the reasoning process. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS conjunction fallacy; probability estimation; rationality; biases

## INTRODUCTION

The capacity to reason with uncertain knowledge (that is, to reason with probabilities) is central to human intelligence. But how do people estimate and reason about probabilities? One revealing aspect of human probabilistic reasoning is the reliable occurrence of the conjunction fallacy in people's probability judgements. Probability theory requires that the probability of a conjunction of two events, $P(A \wedge B)$ (the chances of both *A* and *B* occurring together), can never be greater than the probability of a constituent event *A* (or *B*). This requirement follows from the fact that $A \wedge B$ can only occur if *A* itself occurs. The conjunction fallacy occurs because people reliably violate this requirement for some events, giving probability estimates for conjunctions that are greater than the estimates they gave for a constituent of that conjunction (contrary to the requirements of probability theory).

Two well-known examples of the conjunction fallacy comes from Tversky & Kahneman (1983) and concern Linda and Bill:

> *"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations"*
>
> *"Bill is 34 years old. He is intelligent, but unimaginative, compulsive and generally lifeless. In school, he was strong in mathematics but weak in social studies and humanities."*

Participants in Tversky and Kahneman's study read these descriptions and were asked to rank various statements 'by their probability'. For the Linda description, two of these statements were

$$\textit{Linda is a bank teller.} \qquad (A)$$

$$\textit{Linda is a bank teller and active in the feminist movement.} \qquad (A \wedge B)$$

while for the Bill description, two statements were

$$\textit{Bill plays jazz for a hobby.} \qquad (A)$$

$$\textit{Bill is an accountant who plays jazz for a hobby.} \qquad (A \wedge B)$$

Tversky and Kahneman found that for both of these descriptions, more than 85% of participants ranked $A \wedge B$ as more probable than *A*, committing the conjunction fallacy in a very reliable way. The reliability and robustness of the conjunction fallacy have been confirmed in many subsequent studies (Costello, 2009a; Fantino, Kulik, Stolarz-Fantino, & Wright, 1997; Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991; Sides, Osherson, Bonini, & Viale, 2002; Stolarz-Fantino, Fantino, Zizzo, & Wen, 2003; Tentori, Bonini, & Osherson, 2004; Wedell & Moro, 2008).

While the conjunction fallacy is an undeniable part of human probabilistic reasoning, it does not occur at the same rate for all conjunctions. The conjunction fallacy occurs very frequently for the Linda example (which, after all, is explicitly designed to demonstrate the fallacy) but occurs significantly less frequently for other materials. Numerous experimental studies have shown that the occurrence of this fallacy depends on the probabilities of *A* and *B*. In particular, the lower the probability of the less probable constituent *P(A)*, and the higher both *P(B)* and the conditional probability $P(A|B)$, the more frequent the conjunction fallacy is (Carlson & Yates, 1989; Costello, 2009b; Fantino *et al.,* 1997; Fisk & Pidgeon, 1996; Gavanski & Roskos-Ewoldsen, 1991). For example, Fisk & Pidgeon (1996) found that while the conjunction fallacy occurs at rates up to 85% in cases where *P(A)* was low and *P(B)* was high, the fallacy occurred at rates as low as 10% in cases where *P(A)* and *P(B)* were both low

*Correspondence to: Fintan Costello, School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland. E-mail: fintan.costello@ucd.ie (conjunction fallacy rates from Fisk and Pidgeon (1996) are shown in Tables 2 and 3).

Tversky and Kahneman's demonstration of the conjunction fallacy has had a major impact on our view of the processes behind people's probabilistic reasoning. In particular, the conjunction fallacy contributed to a shift away from the idea that people are 'intuitive statisticians' who approximately follow the tenets of probability theory in their judgements (Peterson & Beach, 1967) to the view that people's probability judgements follow 'a limited number of heuristics which sometimes yield reasonable judgements and sometimes lead to severe and systematic errors' (Tversky & Kahneman, 1973, p. 237). An important aim in research on the psychology of probabilistic judgement, therefore, is to produce a theory explaining the occurrence of the conjunction fallacy. Current theories, however, struggle to explain the wide variation in rates of occurrence of the fallacy for different conjunctions. Our aim in this paper is to show how this wide variation in fallacy rates can be explained in the framework of our 'probability theory plus noise' model. This model rejects the idea that people estimate probabilities using heuristics and instead assumes that people reason according to probability theory but are subject to random noise in the reasoning process; in the model, this random noise causes the systematic biases seen in people's probability estimates (see Costello & Watts, 2014 for details).

The organization of the paper is as follows. In the first section, we provide some background by briefly reviewing current theories of the conjunction fallacy. In the second section, we present our model and show how it can explain the wide variation in conjunction fallacy rates seen in experimental studies. In the three subsequent sections, we test the predictions of this model by applying it to three sets of experimental data from Zhao, Shah, and Osherson (2009), Fisk and Pidgeon (1996) and Costello and Watts (2014). In the final section, we give a general discussion of our results.

## THEORIES OF THE CONJUNCTION FALLACY

Tversky and Kahneman explained the conjunction fallacy in terms of a representativeness heuristic, in which probability is assessed in terms of the degree to which an instance is representative of a (single or conjunctive) category. In this view, people gave a higher rating to the conjunctive statement in the Linda example because the instance Linda was more representative of (that is, more similar to members of) the conjunctive category 'bank-teller and active feminist' than the single category 'bank-teller'.

Although the representativeness heuristic remains the standard explanation of the conjunction fallacy in textbooks, a range of experimental results give convincing evidence against this account. Notice that the representativeness heuristic only applies when a question asks about the probability of membership of an instance in a conjunctive category and when knowledge about representative members of that category is available. Evidence against representativeness comes from results showing that the conjunction fallacy occurs frequently when these requirements do not hold. For example, studies by Osherson, Bonini and colleagues have shown that the conjunction fallacy occurs frequently when people are asked to bet on the occurrence of unique future events: such bets are not questions about membership of an instance in a category, and so representativeness cannot explain the occurrence of the fallacy in these cases (Bonini, Tentori, & Osherson, 2004; Sides *et al.,* 2002; Tentori *et al.,* 2004).

Further problematic results for the representativeness account come from studies by Gavanski and Roskos-Ewoldsen (1991). One study compared fallacy rates for conjunctions for which compelling representativeness information was available (questions such as Tversky and Kahneman's Linda example) against fallacy rates for conjunctions where there was no representativeness information available (questions about the number of eyes of, and hair colour of, aliens on imaginary planets about which no information was available). Materials were constructed so that the probabilities of constituent events were matched across 'representative' and 'non-representative' conjunctions. Gavanski and Roskos-Ewoldsen (1991) found that the conjunction fallacy was just as frequent for conjunctions for which no representativeness information was available as it was for Linda type conjunctions. Indeed, pairing representative and non-representative conjunctions with the same constituent probabilities, they found no significant difference in fallacy rates between these two types of conjunction.

Another study by Gavanski and Roskos-Ewoldsen (1991) provides a different line of evidence against the representativeness account. Notice that the representativeness account implicitly assumes that estimating a conjunctive probability $P(A \wedge B)$ is equivalent to judging the likelihood that an item (the individual Linda) is a member of a conjunctive category $A \wedge B$ (the category of 'people who are bank tellers and active feminists'). While some conjunctive probabilities involve such judgements about an item's membership in a conjunctive category, other conjunctive probabilities do not. Because the representativeness explanation for the conjunction fallacy depends on membership in a conjunctive category, this account predicts that the fallacy will not occur for these other conjunctions.

To test this, Gavanski and Roskos-Ewoldsen (1991) compared fallacy rates in 'standard' and 'mixed' problems using descriptions from Tversky and Kahneman (1983). For standard problems, each conjunction was composed of two events from the same description (conjunctions such as 'Linda is a bank teller and active in the feminist movement' or 'Bill is an accountant who plays jazz for a hobby'). These standard problems thus represented questions about Linda or Bill's membership in conjunctive categories. Mixed problems contained the same target descriptions and component questions as standard problems. However, each conjunction was composed of one event from the first description and one from the second (conjunctions such as 'Linda is active in the feminist movement and Bill plays jazz for a hobby'). These mixed problems thus do not represent questions about Linda or Bill's membership in conjunctive categories. The representativeness account predicts that the conjunction fallacy will occur for the standard problems but not for the

DOI: 10.1002/bdm

mixed problems. Gavanski and Roskos-Ewoldsen (1991) found, however, that the fallacy occurred just as frequently in these two different types of problem, and the occurrence of the fallacy depended only on the probability of the two constituent events in a conjunction. This undermines the idea that the conjunction fallacy occurs because an item is a representative member of a conjunctive category.

This observation also causes problems for other, more recent, accounts in which the conjunction fallacy depends in some way on membership in conjunctive categories or classes. One such is the ‘fuzzy-trace’ or ‘denominator neglect’ account by Reyna et al. (Reyna, 2004; Reyna & Brainerd, 2008; Wolfe & Reyna, 2010). The core of this account is that the conjunction fallacy is a class-inclusion error where people judge an item (Linda) to be a member of a subset class (‘bank tellers that are feminists’) but not a member of the including class (‘bank tellers’). In this account, overlapping class-inclusion relationships create processing interference or confusion and lead people to commit class-inclusion errors such as the conjunction fallacy. The mixed problems used in Gavanski and Roskos-Ewoldsen (1991), however, do not involve class inclusion. Instead, they ask about the probability of Linda being a member of a single class ‘active feminists’ versus the conjunctive probability of Linda being a member of that class and a different individual (Bill) being a member of a different single class (‘people who play jazz as a hobby’). The fact that the conjunction fallacy occurs just as frequently for these mixed problems as for standard conjunction fallacy problems is hard to explain in this account.

If the conjunction fallacy is not caused by differences in representativeness or class-inclusion relationships, what does cause this fallacy? A number of accounts have suggested that the fallacy arises because people compute conjunctive probabilities $P(A \wedge B)$ from constituent probabilities $P(A)$ and $P(B)$ using some form of average of the two constituent probabilities (Carlson & Yates, 1989; Fantino *et al.,* 1997). Most recently, Nilsson, Juslin and colleagues (Juslin, Nilsson, & Winman, 2009; Nilsson, Winman, Juslin, & Hansson, 2009) have proposed a more sophisticated ‘configural weighted average’ model where conjunctive probabilities are computed by a weighted average of constituent probability values, with a greater weight given to the lower constituent probability. Taking $P(A)$ to be the lower of the two constituent probabilities, in this account we have

$$P(A \wedge B) = \beta P(A) + (1 - \beta)P(B)$$

where $\beta$ is a weighting parameter that falls between 0.5 and 1.

All these averaging accounts impose the requirement that $P(A \wedge B) > P(A)$, where $P(A)$ is the lower constituent probability, for all events $A$ and $B$ (except when $P(A) = P(B)$). These accounts thus predict that the conjunction fallacy should occur 100% of the time for all pairs of events (except when $P(A) = P(B)$). This is clearly not the case, however. To address this problem, the Nilsson et al. model also includes a noise component that randomly perturbs individual conjunctive probability estimates, sometimes moving the conjunctive probability below the lower constituent probability and so eliminating the conjunction fallacy for that estimate. Because $P(A \wedge B) > P(A)$, and because this noise component is random (and so moves the conjunctive probability $P(A \wedge B)$ downward 50% of the time and upward the remaining 50%), this configural weighted averaging model predicts that the conjunction fallacy rate should occur at least 50% of the time for all conjunctions. The low conjunction fallacy rates we see for some conjunctions (less than 10%, as in Fisk and Pidgeon (1996)) are thus clearly a challenge for this model.

Finally, in our original ‘probability theory plus noise’ model, the conjunction fallacy occurs purely as a result of random noise in a reasoning process that follows the requirements of probability theory (Costello, 2009b; Costello, 2009a; Costello & Watts, 2014). This model, however, imposes the requirement that $P(A \wedge B) \leq P(A)$ on average, where $P(A)$ is the lower constituent probability. Because on average $P(A \wedge B) \leq P(A)$, and since this noise component is random (and so moves the conjunctive probability $P(A \wedge B)$ downward 50% of the time and upward the remaining 50%), this model predicts that the conjunction fallacy rate should always be 50% or less for all conjunctions. The high conjunction fallacy rates we see for some conjunctions (80% or more in Tversky and Kahneman’s studies and in Fisk and Pidgeon (1996)) are thus clearly a challenge for this probability theory plus noise model. Our aim in this paper is to resolve this challenge by showing how a slight extension of this model gives a natural explanation for these high fallacy rates.

## THE ‘PROBABILITY THEORY PLUS NOISE’ MODEL

Our model assumes that people’s probability judgements are produced by a mechanism that is fundamentally rational but is perturbed in various ways purely random noise or error. In taking this approach, we are following a line leading back at least to Thurstone (1927) and taken up in various ways by other more recent researchers (e.g. Dougherty, Gettys, & Ogden, 1999; Erev, Wallsten, & Budescu, 1994; Hilbert, 2012). Our contribution in this paper is to show that the ‘probability theory plus noise’ account can explain the occurrence of the conjunction fallacy in people’s probability judgements and can explain the wide variation in rates at which that fallacy occurs.

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. In probability theory, the expected value of these estimates is equal to $P(A)$, the probability of $A$; individual estimates vary with an approximately normal distribution around this value. We assume that people estimate the probability of $A$ in exactly this way: randomly sampling episodes from memory, counting the number that are $A$ and dividing by the sample size. We assume a long-term memory from which a random sample of episodes or traces can be drawn. For some event $A$, we assume that each episode $i$ in our sample holds a flag that is set to 1 if $i$ contains event $A$ and set

DOI: 10.1002/bdm

to 0 otherwise. An estimate for the probability of $A$, in this model, is obtained by randomly sampling episodes from memory, counting the number where the flag for $A$ is set to 1 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 a minimal form of transient random noise in which there is some small probability $d < 0.5$ that when some flag is read, the value obtained is not the correct value for that flag. We assume that this noise is symmetric, so that the probability of 1 being read as 0 is the same as the probability of 0 being read as 1. We also assume a minimal representation where every type of event, be it a simple event $A$, a conjunctive event $A \wedge B$, a disjunctive event $A \vee B$ or any other more complex form, is represented by such a flag, and where every flag has the same probability $d$ of being read incorrectly.

A randomly sampled event will be counted as $A$ if the event truly is $A$ and its flag is read correctly (this occurs with a probability *(1 − d)P(A)*, because *P(A)* events are truly $A$ and flags have a *1 − d* chance of being read correctly), or if the event is truly $\neg A$ (not $A$) and its flag is read incorrectly as $A$ (this occurs with a probability *(1 − P(A))d*, because *1 − P(A)* events are truly $\neg A$, and flags have a $d$ chance of being read incorrectly). The expected value for a noisy estimate for the probability of $A$ is thus

$$\begin{aligned}\langle P_E(A)\rangle &= (1-d)P(A) + (1-P(A))d \\ &= (1-2d)P(A) + d \qquad (1)\end{aligned}$$

with individual estimates varying independently with an approximately normal distribution around this value. This average is 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 $P(A) > 0.5$. This model explains a number of observed patterns of bias in people's probability estimates, such as conservatism or underconfidence, subadditivity and binary complementarity (Costello & Watts, 2014). Note that because this model assumes that a probability estimate for $A$ is obtained by counting the proportion of $A$'s in a random sample of events, probability estimates produced by the model necessarily fall in the range $0 \ldots 1$ (as long as the error rate $d$ is a true rate: that is, as long as $d$ also falls in the range $0 \ldots 1$).

Because we assume that the chance of a random error $d$ is the same for conjunctions and disjunctions of events as it is for individual events, we also obtain

$$\langle P_E(A \wedge B)\rangle = (1-2d)P(A \wedge B) + d \qquad (2)$$

and

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

From this, the model also makes the surprising prediction that while people's probability estimates will be biased away from the 'true' value in various ways, for some specific expressions, this bias will 'cancel out'; for these expressions, people's probability estimates will agree with the requirements of probability theory. One such cancelling expression is probability theory's 'addition law':

$$P_E(A) + P_E(B) - P_E(A \wedge B) - P_E(A \vee B) = 0$$

Probability theory requires that this expression has a value of 0 for all events $A, B$. Substituting the model's predicted expected values into this expression, we see that the model predicts an average value of 0 for this expression, just as required by standard probability theory. A series of experiments show that this prediction does in fact hold: values for this expression, computed from people's individual probability estimates, are on average strikingly close to zero, just as required (Costello & Watts, 2014). (Note that this value of 0 for the addition law is also predicted by other accounts of the conjunction fallacy, such as the configural weighted averaging account of Nilsson and colleagues; see Nilsson, Juslin, and Winman (2016) for this prediction, and Costello and Watts (2016) for our response. We also address the distinction between these two models in our analysis of experimental values for the addition law expression subsequently).

The conjunction fallacy occurs in this model because individual estimates for both *P(A)* and $P(A \wedge B)$ vary randomly around their expected values. This random variation means that some individual estimates will occur where $P_E(A) < P_E(A \wedge B)$, producing a conjunction fallacy response. The closer the expected values $\langle P_E(A)\rangle$ and $\langle P_E(A \wedge B)\rangle$ are to each other, the greater the chance of this fallacy response occurring. These values are closest when *P(A)* is low and both *P(B)* and $P(A|B)$ are high, and so the model predicts that the conjunction fallacy will be most frequent when these criteria hold. This is just the pattern seen in studies of the conjunction fallacy (Carlson & Yates, 1989; Costello, 2009b; Fantino *et al.,* 1997; Gavanski & Roskos-Ewoldsen, 1991).

This model is also consistent with more specific results showing that the degree to which $B$ causes $A$ has an impact on conjunction fallacy rates (Pidgeon, 1998; Tentori, Crupi, & Russo, 2013; Tversky & Kahneman, 1983). We illustrate this using Tversky and Kahneman's original experiment investigating the role of causality on conjunction fallacy occurrence. If an event $B$ causes $A$, then by definition $P(A|B) > P(A)$ (the probability of $A$, given the causing event $B$ is higher than the probability of $A$ simpliciter). Tversky & Kahneman (1983) gave one group of participants a problem where this causal link held; that is, where $P(A|B) > P(A)$:

*A health survey was conducted in a representative sample of adult males in British Columbia of all ages and occupations. Mr. F. was included in the sample. He was selected by chance from the list of participants.*

*Which of the following statements is more probable? (check one)*

- *Mr. F. has had one or more heart attacks.* $(A)$
- *Mr.F. has had one or more heart attacks and he is over 55 years old.* $(A \wedge B_1)$

**DOI**: 10.1002/bdm

A substantial proportion (58 %) of participants produced the conjunction fallacy for this problem. Tversky and Kahneman then gave another group of participants an alternative problem:

*A health survey was conducted in a representative sample of adult males in British Columbia of all ages and occupations. Mr. F. and Mr. G. were both included in the sample. They were unrelated and were selected by chance from the list of participants.*

*Which of the following statements is more probable? (check one)*

- *Mr. F. has had one or more heart attacks.* $(A)$

- *Mr. F has had one or more heart attacks and Mr. G. is over 55 years old.* $(A \wedge B_2)$

This alternative has the same constituent probabilities as in the original problem (the event $A$ is the same across both problems, and $B_1$ and $B_2$ refer to the same event: a randomly sampled person being over 55 years). In this alternative problem, however, we have $P(A|B_2) = P(A)$ because the chance of Mr. F. having had a heart attack (event $A$) is not affected by the additional information that Mr. G. is over 55 years old (event $B_2$). A significantly lower proportion (29 %) of participants in Tversky and Kahneman's study produced the conjunction fallacy for this alternative problem, showing that the causal link between constituents influences conjunction fallacy rates.

This pattern of results is just as expected in our model. From Equation (2), we have

$$\begin{aligned}\langle P_E(A \wedge B_1)\rangle &= (1 - 2d)P(A \wedge B_1) + d \\ &= (1 - 2d)P(A|B_1)P(B_1) + d\end{aligned}$$

and

$$\begin{aligned}\langle P_E(A \wedge B_2)\rangle &= (1 - 2d)P(A \wedge B_2) + d \\ &= (1 - 2d)P(A|B_2)P(B_2) + d \\ &= (1 - 2d)P(A)P(B_2) + d\end{aligned}$$

Because these problems are designed so that $P(A|B_1) > P(A)$ and $P(B_2) = P(B_1)$, we see that $\langle P_E(A \wedge B_1)\rangle > \langle P_E(A \wedge B_2)\rangle$ necessarily holds. This in turn implies that $\langle P_E(A \wedge B_1)\rangle$ is closer to $\langle P_E(A)\rangle$, and so the model predicts higher conjunction fallacy rates for $A \wedge B_1$, just as observed.

In this model, the difference between the expected value for *P(A)* and the expected value for $P(A \wedge B)$ is

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

Because by definition, $P(A) - P(A \wedge B) \geq 0$, and by assumption, $d < 0.5$, this model predicts that

$$\langle P_E(A \wedge B)\rangle \leq \langle P_E(A)\rangle$$

must always hold: that is, with the same rate of error $d$ for single events and conjunctions, this model requires the average value for the conjunctive probability $P_E(A \wedge B)$ will never be greater than the average value for its constituent $A$. Since individual estimates $P_E(A \wedge B)$ and $P_E(A)$ are both perturbed by random noise (which is equally likely to be positive or negative), the requirement that $\langle P_E(A \wedge B)\rangle \leq P_E(A)\rangle$ means that an individual estimate $P_E(A \wedge B)$ can randomly fall above an estimate $P_E(A)$ at most 50 % of the time. While fallacy rates of less than 50 % do hold for many events, there are some pairs of events for which fallacy rates are significantly higher than this bound (Tversky and Kahneman's Linda and Bill being two examples). We now show how a slight extension of the model can explain these high fallacy rates, by allowing the average value for the conjunctive probability to be greater than the average value for its constituent in certain situations.

### Increased noise in combined expressions

In our original presentation of the probability theory plus noise model, as described earlier, we assumed that the occurrence of all events (simple events and any form of conjunction or disjunction) is recorded via flags representing those events. We also assumed that the rate of random error $d$ is the same for all flags, and so there is the same rate of error when counting instances of a simple event $A$ and when counting instances of a conjunction $A \wedge B$ or disjunction $A \vee B$. These assumptions give a useful first-order approximation representing a noisy memory system, but are clearly unrealistic: they assume, for example, that all possible conjunctions or disjunctions that could ever be observed are explicitly represented by flags in memory. Here, we give a slightly more realistic version of this model: a second-order approximation of noisy memory.

As before, the probability of some event $A$ is obtained by sampling a set of episodes and counting the number that contains $A$. Rather than assuming that the presence or absence of $A$ in a given episode $i$ is represented by a flag, we assume that there is some classification mechanism that takes episode $i$ as input and processes it to give a decision indicating whether the episode is an example of $A$ or not. We assume that this process is subject to symmetric random noise such that for a given event $A$, it has a chance $d$ of mistakenly classifying items that are truly $A$ as not $A$ and the same chance of mistakenly classifying items that are truly not $A$ as $A$. Finally, we assume that the level of random noise $d$ associated with the classification of episodes as examples of $A$ depends in some way on the complexity of $A$: the more complex $A$ is, the more difficult the classification process is, and so the greater the value of $d$. We do not give any formal description of the complexity of different types of events (we address this informally in the Discussion section below), beyond noting that conjunctions $A \wedge B$ and disjunctions $A \vee B$ are more complex than their constituent events $A$ and $B$.

Our original model assumed that the error rate $d$ was the same for all types of event. In this extended model, however, we would expect the rate of random error when counting conjunctions and disjunctions to be slightly higher than the rate when counting occurrences of single events. This is

DOI: 10.1002/bdm

because classifying a given instance as an example of a conjunction $A \wedge B$ or a disjunction $A \vee B$ is more complex than classifying it as an example of one of the constituent categories $A$ and $B$ (classification in a conjunction $A \wedge B$ or a disjunction $A \vee B$ requires two component decisions, one for each constituent, while classification in $A$ or $B$ requires only a single such decision). To reflect this expectation of increased random error in counting conjunctions and disjunctions, we assume a rate of random error of $d$ for single events but of $d + \Delta d$ for conjunctions and disjunctions (where $\Delta d$ represents some small increase in the rate of random error). 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 extended model, we have $\langle P_E(A)\rangle = (1 - 2d)P(A) + d$ as before, but

$$\langle P_E(A \wedge B)\rangle = (1 - 2[d + \Delta d])P(A \wedge B) + [d + \Delta d] \qquad (3)$$

and

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

With this extension, we see that the model no longer requires $\langle P_E(A \wedge B)\rangle \leq \langle P_E(A)\rangle$; instead $\langle P_E(A \wedge B)\rangle > P_E(A)\rangle$ can hold when

$$(1 - 2[d + \Delta d])P(A \wedge B) + [d + \Delta d] > (1 - 2d)P(A) + d$$

or equivalently when

$$\Delta d[1 - 2P(A \wedge B)] > (1 - 2d)[P(A) - P(A \wedge B)] \qquad (4)$$

From this expression, we see that the average estimate for $A \wedge B$ can be higher than the average estimate for its constituent $A$ when $\Delta d$ is positive; when $1 - 2P(A \wedge B) > 0$ (that is, when $P(A \wedge B) < 0.5$); and when $P(A) - P(A \wedge B)$ is small (that is, when $P(A \wedge B)$ is close to $P(A)$). When these three requirements hold, the model predicts the occurrence of the conjunction fallacy in averaged estimates; that is, it predicts $\langle P_E(A \wedge B)\rangle > \langle P_E(A)\rangle$. Because individual estimates $P_E(A \wedge B)$ and $P_E(A)$ are both perturbed by random noise (which is equally likely to be positive or negative), when these requirements hold and so $\langle P_E(A \wedge B)\rangle > \langle P_E(A)\rangle$, we expect that an individual estimate $P_E(A \wedge B)$ will randomly fall above an estimate $P_E(A)$ more than 50% of the time. In other words, this model predicts that the conjunction fallacy should occur in individual estimates *more than* 50% of the time when these requirements hold. The greater the difference between the left and right-hand terms in this inequality, the greater the rate of conjunction fallacy occurrence.

This modification leads to a slight change in the model's predictions about the cancelling expressions described earlier. For the addition law, we obtain an expected value of

$$\begin{aligned}(1 - 2d)P(A) + d + (1 - 2d)P(B) + d \\ -(1 - 2(d + \Delta d))P(A \wedge B) - (d + \Delta d) \\ -(1 - 2(d + \Delta d))P(A \vee B) - (d + \Delta d) \\ = 2\Delta d(P(A \wedge B) + P(A \vee B) - 1) \\ = 2\Delta d(P(A) + P(B) - 1)\end{aligned}$$

and because $P(A) + P(B) - 1$ takes only values in the range $-1$ to $+1$, the expected value for this expression is predicted to always be within $2\Delta d$ of 0. In other words, while the original model predicted that values for the addition law would be distributed around a mean of 0 for all pairs of events $A, B$, this model predicts that for each pair $A, B$, these values will be distributed around a mean that is close to 0 (within $2\Delta d$ of 0), with the precise value of that mean varying with the probabilities *P(A)* and *P(B)*. Because in this model, we also have

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

and

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

(by rearranging Equations (1) and (3)), we expect that values for the addition law will vary around zero in a way that is directly proportional to both $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$ and $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$. As far as we can see, Nilsson and colleagues' configural weighted averaging model (Nilsson *et al.,* 2009; Nilsson *et al.,* 2016) does not make these predictions, allowing us to distinguish between the two models. We investigate these predictions in the section on the addition law below.

Just like the original model, this extended model explains observed patterns of bias in people's probability estimates such as conservatism or underconfidence, subadditivity and binary complementarity (Costello & Watts, 2014). This is because all those patterns involve single events only and not conjunctions or disjunctions. In the next sections, we test this extended model in three ways: by applying to probability estimation data from Zhao *et al.* (2009), to conjunction fallacy rates from a study by Fisk and Pidgeon (1996) and to data about variation in single and conjunctive probability estimates and in values of the addition law from a previous experiment of our own (Experiment 1 in Costello (2009a)). In each case, the results support the proposed model.

## VARIATION AND VALUES OF $d$ AND $\Delta d$

How realistic is the idea that noise rates are higher in conjunctive probability estimates than in constituent probability estimates? We can test this idea in two ways: by comparing the standard deviations (SDs) of conjunctive and constituent probability estimates (if noise rates are higher for conjunctions, we would expect higher SDs) and by considering the degree to which people's probability estimates for conjunctive and single events differ from the true objective

DOI: 10.1002/bdm

probabilities of those events (if noise rates were higher for conjunctions, we would expect this difference between estimated and true probabilities to be greater for conjunctions). We can carry out both these tests using results reported by Zhao *et al.* (2009).

Zhao *et al.* (2009) describe two experiments where participants viewed 12 sets of geometric shapes on a computer screen. Each set was a mixture of 20 triangles, squares and circles in blue, red and green (all three shapes and all three colors appeared in every set). For each set, Zhao *et al.* (2009) chose one colour and one shape to serve as the categories A and B. A different choice was made for each of the 12 sets; for six sets, *A* was a colour and *B* a shape, the reverse held for the other six. After viewing each set, participants were asked to estimate the probability that a randomly selected item from that set would be *B* or the probability that a randomly selected item from that set would be $A \wedge B$. If *A* was 'red' and *B* was 'square' in a given set, for example, participants were asked to estimate the probability that a random item from that set would be red or the probability that a random item from that set would be a red square. (Participants were also asked to estimate conditional probabilities $P(A|B)$ and $P(B|A)$; however, we do not consider these estimates in our analysis). There were 45 participants in each experiment: Zhao *et al.* (2009) averaged across all these participants estimates for low, medium and high sets (Table 1).

As Table 1 shows, the SDs for low, medium and high conjunctive probabilities $A \wedge B$ were higher than the SDs for the corresponding low, medium and high constituent probabilities *B* in both experiments. This is consistent with the proposal that random variation is higher in conjunctions than in single events. Table 1 also shows that the difference between objective probabilities and probability estimates was higher for conjunctions than for single events (root mean-squared differences of 0.08 versus 0.03 in both experiments). This is also consistent with our proposal, where the difference between objective and estimated probabilities would be higher for conjunctions because of the assumed higher noise rates. As a further test of this proposal, we considered whether the absolute difference between objective and estimated probabilities was related to the SD of those estimated probabilities. Recall that in the probability theory plus noise model, increased noise (and so increased SD in estimated probabilities) should be related to increased difference between estimated and objective probabilities. Across the 12 probability estimates in Table 1, there was a reliable positive correlation between degree of difference between estimated and objective probabilities, and the SD of estimated probabilities ($r(10) = 0.62, p < 0.05$). This is in line with the model's predictions.

Finally, as an additional test, we fitted the probability theory plus noise equations for single and conjunctive probability estimates (Equations (1) and (3)) to the average estimates given by participants in the experiments (Table 1). To perform this fit, we simply took the objective probabilities used in the experiment and searched for the values of $d$ and $\Delta d$ which, when combined with those objective probabilities as in the equations, produced probability estimates that were closest to the average probability estimates given in the experiment (in terms of root mean-squared differences). In both experiments, the best fitting values were $d = 0.05$ and $\Delta d = 0.04$, supporting the idea of greater random noise in conjunctive estimates than constituents.

## SIMULATION OF THE MODEL

It is clear from Equation (4) that the conjunction fallacy rate in this model is related to the values of $P(A)$, $P(A \wedge B)$ and values of $d$ and $\Delta d$. We examined this relationship in detail via Monte Carlo simulation, by writing a program that simulates the effects of random noise in recall on probability estimations for a given set of probabilities. This 'single-estimate' simulation program took as input three probabilities $P_I(A)$, $P_I(B)$ and $P_I(A \wedge B)$. The program constructed a 'memory' containing 250 items, 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

Table 1. Objective probability values and average probability estimates (and SD) for $B$ and $A \wedge B$ from Tables 2 and 3 in Zhao *et al.* (2009), along with best-fitting estimates from Equations (1) and (3) and RMSD between participants' average probability estimates and objective probability values

| | Objective probability | Estimates (Experiment 1) | | Estimates (Experiment 2) | |
|---|---|---|---|---|---|
| | | Participants (SD) | Model | Participants (SD) | Model |
| *P(B)* | 0.3 | 0.31 (0.07) | 0.32 | 0.29 (0.05) | 0.32 |
| *P(B)* | 0.6 | 0.58 (0.07) | 0.59 | 0.59 (0.08) | 0.59 |
| *P(B)* | 0.9 | 0.86 (0.04) | 0.86 | 0.85 (0.05) | 0.86 |
| RMSD from objective probabilities | | 0.03 | | 0.03 | |
| $P(A \wedge B)$ | 0.1 | 0.19 (0.08) | 0.17 | 0.17 (0.06) | 0.17 |
| $P(A \wedge B)$ | 0.4 | 0.5 (0.10) | 0.42 | 0.51 (0.16) | 0.42 |
| $P(A \wedge B)$ | 0.8 | 0.8 (0.07) | 0.75 | 0.78 (0.06) | 0.75 |
| RMSD from objective probabilities | | 0.08 | | 0.08 | |

SD, standard deviation; RMSD, root mean-squared differences.

SDs were higher for conjunctive estimates than single estimates. RMSD values were higher for conjunctions, indicating that the difference between estimates and objective probabilities was higher for conjunctions. For both experiments, the best-fitting results were produced by the model equations with $d = 0.05$ and $\Delta d = 0.04$.

**DOI**: 10.1002/bdm

noise parameter values $d$ and $\Delta d$, representing the noise rate and the increase in noise for conjunctions. When reading flag values from memory to generate some probability estimate for a single event $P_E(A)$, the program was designed to have a random chance $d$ of returning the incorrect value for a flag; when reading flags to generate an estimate for a combined event $P_E(A \wedge B)$, the program had a random chance equal to $d + \Delta d$ of returning the incorrect value for a flag. To produce an estimate for the probability of some given event, the program simply went through the memory reading the values of flags for that event (subject to this random error in reading) and returned the proportion of flags that were read as true as its estimate for the probability of the given event.

For a given set of input probabilities $P_I(A)$, $P_I(B)$ and $P_I(A \wedge B)$, each run of this single-estimate simulation program generated a single noisy estimate for each of the probabilities $P_E(A)$, $P_E(B)$ and $P_E(A \wedge B)$. These represent a single individual's estimates for the probabilities of these events. A conjunction fallacy occurs in these estimates if $P_E(A) < P_E(A \wedge B)$ or $P_E(B) < P_E(A \wedge B)$.

We used this single-estimate simulation program to test the extent to which the probability theory plus noise model can match the range of conjunction fallacy rates observed in experimental studies. To carry out this test, we applied the simulation to the conjunction fallacy results reported in an experiment by Fisk and Pidgeon (1996). That experiment used a within-subjects design with the type of conjunction being the principal independent variable. Three types of $A \wedge B$ conjunction were constructed: LL (where $A$ and $B$ were both likely); LU (where $A$ was likely and $B$ unlikely) and UU (where both $A$ and $B$ were unlikely). Participants in the experiment saw questions about 10 different scenarios, including the Linda scenario, other scenarios from Tversky and Kahneman (1983) and various similar scenarios. For each scenario, participants gave probability judgements for various constituents $A$, $B$ and for the likelihood of the three types of conjunction $A \wedge B$ (the LL conjunctions, LU conjunctions and UU conjunctions). Fisk and Pidgeon (1996) report the average probability estimate for each constituent and each conjunction (in their Table 5). For each scenario and conjunction type, they also report single conjunction fallacy rates (where participants judge a conjunction to be more likely than one of its constituents), double conjunction fallacy rates (where participants judge a conjunction to be more likely than both of its constituents) and non-fallacy rates (in their Table 2).

To apply this single-estimate simulation program to Fisk and Pidgeon's data, we created a 'multiple-estimate'

Table 2. Observed and simulated probability estimates and single conjunction fallacy rates

| | Conjunction | $\langle P_E(A)\rangle$ | $\langle P_E(B)\rangle$ | $\langle P_E(A \wedge B)\rangle$ | | Single fallacy rate | |
|---|---|---|---|---|---|---|---|
| Scenario | type | (Observed) | (Observed) | (Observed) | (Simulation) | (Observed) | (Simulation) |
| 1 | LL | 0.84 | 0.62 | 0.71 | 0.61 | 51.1 | 39.1 |
| 1 | LU | 0.84 | 0.21 | 0.42 | 0.22 | 83.9 | 65.7 |
| 1 | UU | 0.32 | 0.14 | 0.14 | 0.13 | 31.1 | 31.3 |
| 2 | LL | 0.66 | 0.76 | 0.73 | 0.65 | 33.3 | 36.6 |
| 2 | LU | 0.66 | 0.19 | 0.31 | 0.20 | 69.2 | 68.8 |
| 3 | LL | 0.66 | 0.76 | 0.74 | 0.65 | 35.5 | 35.6 |
| 3 | LU | 0.76 | 0.28 | 0.37 | 0.28 | 64.8 | 60.2 |
| 3 | UU | 0.13 | 0.14 | 0.09 | 0.14 | 31.8 | 32.2 |
| 4 | LL | 0.74 | 0.76 | 0.71 | 0.73 | 31.2 | 26.5 |
| 4 | LU | 0.74 | 0.26 | 0.35 | 0.27 | 59.3 | 62.3 |
| 4 | UU | 0.18 | 0.26 | 0.13 | 0.18 | 39.6 | 39.9 |
| 5 | LL | 0.69 | 0.69 | 0.66 | 0.68 | 35.9 | 30.4 |
| 5 | LU | 0.69 | 0.33 | 0.39 | 0.33 | 48.4 | 50.4 |
| 5 | UU | 0.33 | 0.11 | 0.13 | 0.11 | 53.8 | 57.7 |
| 6 | LL | 0.68 | 0.77 | 0.69 | 0.67 | 42.7 | 37.5 |
| 6 | LU | 0.68 | 0.17 | 0.21 | 0.17 | 50.0 | 47.1 |
| 6 | UU | 0.24 | 0.17 | 0.08 | 0.14 | 6.6 | 5.3 |
| 7 | LL | 0.74 | 0.70 | 0.61 | 0.69 | 27.2 | 28.0 |
| 7 | LU | 0.74 | 0.26 | 0.32 | 0.26 | 55.1 | 52.5 |
| 7 | UU | 0.27 | 0.26 | 0.19 | 0.27 | 31.1 | 34.0 |
| 8 | LL | 0.82 | 0.84 | 0.79 | 0.80 | 18.4 | 16.3 |
| 8 | LU | 0.82 | 0.23 | 0.31 | 0.23 | 52.4 | 52.4 |
| 8 | UU | 0.33 | 0.18 | 0.11 | 0.17 | 20.9 | 20.1 |
| 9 | LL | 0.59 | 0.85 | 0.63 | 0.58 | 42.9 | 40.3 |
| 9 | LU | 0.85 | 0.31 | 0.42 | 0.32 | 69.0 | 56.3 |
| 9 | UU | 0.25 | 0.13 | 0.09 | 0.12 | 26.4 | 25.6 |
| 10 | LL | 0.70 | 0.78 | 0.71 | 0.69 | 47.1 | 32.7 |
| 10 | LU | 0.78 | 0.18 | 0.28 | 0.19 | 60.0 | 58.6 |
| 10 | UU | 0.28 | 0.18 | 0.12 | 0.17 | 25.6 | 29.3 |

Observed data are from Fisk and Pidgeon (1996). Simulated data are with $d = 0.05$ and $\Delta d = 0.015$. Simulated estimates for constituent probabilities are not reported, as they exactly matched the observed estimates in all cases. The mean absolute difference between simulated and observed fallacy rates was $M = 3.9$ on the 100-point percentage scale; the mean absolute difference between simulated and observed conjunctive probability estimates was less than 0.06 (comparing across both constituent and conjunctive probability estimates, the mean absolute difference between observed and simulated probability estimates was 0.02). The correlation between observed and simulated fallacy rates was $r = 0.95, p < 0.0001$; the correlation between observed and simulated conjunctive probability estimates was $r = 0.96, p < 0.0001$. Conjunction fallacy rates in the simulation ranged from as high as 68.8 to as low as 5.3.

**DOI**: 10.1002/bdm

simulation program. This program took as input fixed values of $d$ and $\Delta d$, all $\langle P_E(A)\rangle$ and $\langle P_E(B)\rangle$ values from the Fisk and Pidgeon (1996) data and the single and double conjunction fallacy rates for each conjunction. For each conjunction $A$ and $B$, the multiple-estimate program used the given estimates $\langle P_E(A)\rangle$ and $\langle P_E(B)\rangle$ and the fixed value input of $d$ to compute input probabilities $P_I(A)$ and $P_I(B)$ for the single-estimate simulation, using the expression

$$P_I(A) = \frac{\langle P_E(A)\rangle - d}{1 - 2d}$$

This expression is the inverse of our expression for the average value of a probability estimate $\langle P_E(A)\rangle$ given the 'true' probability *P(A)*. By using this expression to compute the input probabilities for our simulation, we ensure that the average of the single-estimate simulation's noisy estimates for the probability of $A$ and $B$ will be the same as the average estimate given by Fisk and Pidgeon's participants. To find an input probability for the conjunction $P_I(A \wedge B)$, the program iteratively considered all possible values of $P_I(A \wedge B)$ that were less than the minimum of $P_I(A)$ and $P_I(B)$ and greater than both 0 and $P_I(A) + P_I(B) - 1$ (and so were consistent with the rules of probability theory). For each of these values, the multiple-estimate program called the single-estimate program 1000 times with inputs $P_I(A)$, $P_I(B)$, $d$ and $\Delta d$. This generated 1000 noisy estimates $P_E(A)$, $P_E(B)$ and $P_E(A \wedge B)$; the program then computed the rate at which single conjunction fallacies occurred in those individual estimates. The value of $P_I(A \wedge B)$ for which this computed single conjunction fallacy rate was closest to the single conjunction fallacy rate observed by Fisk and Pigeon for that conjunction was recorded as the best fitting value of $P_I (A \wedge B)$.

Note that this simulation process applied the same values of $d$ and $\Delta d$ to all probability estimates from the Fisk and Pidgeon data: the values $d$ and $\Delta d$ were the same in each run of the simulation across all conjunctions and scenarios. To find values of $d$ and $\Delta d$ that gave the closest overall agreement with Fisk and Pidgeon's results, we simply ran the overall simulation a number of different times with various different values of $d$ and $\Delta d$.

### Simulation results

Table 2 shows the results of this simulation of Fisk and Pidgeon's data for the best-fitting values of $d$ and $\Delta d$ ($d = 0.05$ and $\Delta d = 0.015$; a similar fit was obtained for a range of other similar values for $d$ and $\Delta d$). As this table shows, the single fallacy rates produced by the simulation for a given pair of constituent probabilities $A$,$B$ closely agreed with the observed single fallacy rates for those conjunctions. The mean absolute difference between observed and simulated fallacy rates was low (=3.9 on the 100-point percentage scale), and the correlation between observed and simulated fallacy rates was high ($r = 0.95, p < 0.0001$). The chance-corrected coefficient of identity between observed and simulated fallacy rates was $e = 0.99$ (where $e = 1.0$ indicates perfect identity; see Zegers and Ten Berge (1985); Zegers (1986) for this measure of identity). The mean absolute difference between observed and simulated conjunctive probability estimates was also low ($M = 0.06$), and their correlation and coefficient of identity values were also high ($r = 0.96, p < 0.0001, e = 0.98$). Simulated probability estimates for single events $A$ and $B$ were identical to the observed probability rates in all cases, giving an overall mean absolute difference between observed and simulated probability estimates (across both single and conjunctive events) of less than 0.02.

There was little difference between the overall average single fallacy rates in the experiment ($M$=42.9 %) and in the simulation ($M$=40.4 %). A similar, although less close, fit was obtained for double conjunction fallacies: the average absolute difference between simulated and observed double fallacy rates was less than 7 %, and overall double fallacy rates were higher in the experiment ($M = 9.1$ %) than in the simulation ($M = 1.5$ %). Combining both single and double fallacy data, the correlation between observed and simulated fallacy rates was very high ($r = 0.93$, $p < 0.0001, e = 0.97$).

Taken together, these results show that the probability theory plus noise model can give a close match to the probability estimates for single events, probability estimates for conjunctive events and conjunction fallacy rates seen in Fisk and Pidgeon (1996). Note, however, that these results represent a best fit, produced when the model is allowed to select the best-matching value for each of the 29 conjunctive input probabilities $P_I(A \wedge B)$.

One possible concern readers might have with our simulation is that the 29 parameters $P_I(A \wedge B)$ representing conjunctive probabilities would allow us to fit any set of 29 data points, and so the fact that the model fits the 29 data points in the Fisk and Pidgeon conjunction fallacy data is not surprising. We think this concern is unfounded because the 29 conjunctive probability parameters are not in fact fully free: these parameters cannot take any value from 0 to 1 but instead are subject to the probability theory constraints that $P_I(A \wedge B)$ is less than the minimum of $P_I(A)$ and $P_I(B)$ and greater than both 0 and $P_I(A) + P_I(B) - 1$. These restrictions mean that parameters $P_I(A \wedge B)$ can only vary within a limited range: recovering values of $P_I(A)$ and $P_I(B)$ from the Fisk and Pidgeon data, we find that the $P_I(A \wedge B)$ parameters can only vary within a range of approximately 0.15. To acquire a measure of the degree of freedom this limited range allows, we can consider Fisk and Pidgeon's 29 data points as corresponding to a single point in 29-dimensional cube of side 1 and volume 1, where each side of the cube represents the value of one of our 29 parameters $P_I(A \wedge B)$. If the $P_I(A \wedge B)$ parameters were fully free, their values could vary between 0 and 1 and some choice of parameters values could fit *any* point in that cube, and so some choice of values would necessarily fit Fisk and Pidgeon's data. The restriction on values of parameters $P_I(A \wedge B)$ that arise because of probability theory's constraints, however, means that these parameters can only fit points that lie within a 29-dimensional subcube with side approximately 0.15 (this is because these parameters cannot vary between 0 and 1; instead, each parameter can only vary within a range of approximately

**DOI**: 10.1002/bdm

0.15). The volume of this 'fittable' subcube is approximately $0.15^{29} \approx 10^{-24}$, and so this subcube represents an absolutely miniscule proportion of the overall space. Even with these 29 $P_I(A \wedge B)$ parameters, in other words, the set of datapoints the model can fit is extremely restricted, and the fact that the model fits the Fisk and Pidgeon data well provides strong support for the model.

A second concern is that these parameters $P_I(A \wedge B)$ may have allowed the simulation to fit the Fisk and Pidgeon conjunction fallacy data in a way that is not consistent with the actual conjunctive probabilities in Fisk and Pidgeon's experiment. As we saw earlier, however, the conjunctive probability estimates produced by the simulation were relatively close to the conjunctive probability estimates produced by participants in Fisk and Pidgeon's experiment: the simulation produced fallacy rates that were similar to those seen by Fisk and Pidgeon while using conjunctive probability estimates that were consistently close to those in Fisk and Pidgeon's study.

Overall, these results thus show that the noise model can produce probability estimates and fallacy rates matching those of Fisk and Pidgeon (1996) while retaining consistency with probability theory in the underlying 'true' input probabilities. These results do not demonstrate that the probability theory plus noise model explicitly predicts the probability estimates and fallacy rates seen by Fisk and Pidgeon. Our aim here is to show that the noise model *can* match those results, not that it predicts those results in detail.

These simulation results demonstrate a number of important aspects of our model. First, they show that the model can produce a wide range of conjunction fallacy rates for different conjunctions: conjunction fallacy rates in the simulation ranged from as high as 68.8 % to as low as 5.3 %. Second, they show that the model can produce probability estimates with varying relationships between the conjunctive probability and the constituent probabilities. In particular, while in many cases, the average conjunctive probability produced by the model was less than both constituent probabilities, in some cases, the average conjunctive probability was higher than one of the constituent probabilities.

To what extent is the observed fit between this simulation and Fisk and Pidgeon's results dependent on the specific values of $d$ and $\Delta d$? To examine this, we ran the simulation again with a range of different values for these parameters ($d$ ranging from 0.01 to 0.20 in steps of 0.01 and $\Delta d$ ranging from 0.01 to 0.10 in steps of 0.01). Figure 1 illustrates the relationship between one measure of goodness-of-fit (mean absolute difference between simulated and observed fallacy rates) and values of $\Delta d$ between 0.01 and 0.1, for five different values of $d$ ($d = 0.01, 0.05, 0.1, 0.15, 0.2$; the same relationship held for all other values of $d$). This figure shows that the degree of fit did not depend greatly on the value of $d$; it did, however, depend on the value of $\Delta d$. In particular, the figure shows that the mean absolute difference in fallacy rates was low (less than 10 on the 100-point percentage scale) for all values of $d$ when $\Delta d = 0.01$ but high (greater than 30) for all values of $d$ when $\Delta d = 0.1$. This graph suggests that the simulation gives results that are fairly consistent with Fisk and Pidgeon's results for all values of $d < 0.1$ and $\Delta d < 0.05$: they do not depend on the specific values of $d$ and $\Delta d$ used in our initial simulation. The same pattern held for our other measure of fit, the coefficient of identity: with $\Delta d = 0.01$, the coefficient of identity was high for all values of $d$ ($M = 0.97$), but with $\Delta d = 0.1$, the coefficient of identity was lower for all values of $d$ ($M = 0.78$). The fact that the degree of fit to conjunction fallacy data depended primarily on the value of $\Delta d$ is in line with our proposal that the observed variation in fallacy rates can be explained as a consequence of increased noise in conjunctive probability estimates (that is, in terms of $\Delta d$).

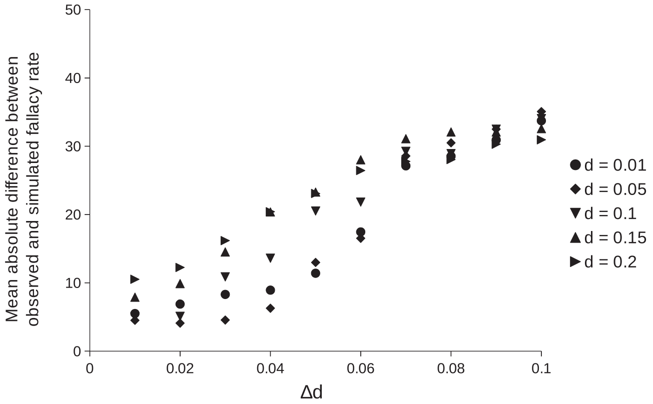

Figure 1. This figure shows how one measure of goodness-of-fit between our simulation and Fisk and Pidgeon's conjunction fallacy data depended on values of $\Delta d$ between 0.01 and 0.1, for five different values of $d$ ($d = 0.01, 0.05, 0.1, 0.15, 0.2$). Goodness-of-fit was measured in terms of mean absolute difference between fallacy rates for events in our simulation and fallacy rates for the same events in Fisk and Pidgeon

## Fallacy rates for different conjunction types

Perhaps the main result from Fisk and Pidgeon's experiment was that conjunction fallacy rates were highest for conjunctions where one constituent was likely and the other was unlikely (LU, or *likely-unlikely*, type conjunctions), next highest for LL conjunctions, and marginally lowest for UU conjunctions (Table 2). Can our model account for this pattern? Recall from our discussion of Equation (4) that conjunction fallacy rates in our model depend on two things: the difference between $P(A \wedge B)$ and the probability of the least likely constituent (the smaller that difference, the greater the chance of a conjunction fallacy relative to that constituent) and the extent to which $P(A \wedge B) < 0.5$: if $P(A \wedge B)$ is close to the probability of the least likely constituent and $P(A \wedge B) < 0.5$, the model predicts fallacy rates of over 50 %, with rates rising with closer conjunctive and constituent probabilities and with lower $P(A \wedge B)$ values. We can use these two factors to make predictions about fallacy rates in LU, LL and UU conjunctions. We consider each conjunction type separately here.

For LU conjunctions, we assume $P(A) > 0.5 > P(B)$ ($A$ is likely and $B$ is unlikely). In this case, the least likely constituent is $B$, and fallacy rates are predicted to be high when the difference between $P(A \wedge B)$ and $P(B)$ is small. For LU conjunctions, however, the difference between $P(A \wedge B)$ and $P(B)$ will always tend to be small and will decline as

DOI: 10.1002/bdm

*P(A)* rises. Why is this? Probability theory requires that $P(A \wedge B) \geq P(A) + P(B) - 1$ must hold for all events, and so $P(A \wedge B)$ must always be within $1 - P(A)$ of *P(B)*. For LU conjunctions, *P(A)* is high and so $1 - P(A)$ is small, and $P(A \wedge B)$ must always be relatively close to *P(B)*, growing closer and closer as *P(A)* rises. For LU conjunctions, we also have $P(A \wedge B) < 0.5$, because $P(A \wedge B) < P(B)$ necessarily holds and $P(B) < 0.5$. Because for LU conjunctions, the difference between $P(A \wedge B)$ and *P(B)* will tend to be small and $P(A \wedge B) < 0.5$, the noise model predicts high fallacy rates for LU conjunctions (fallacy rates tending to be over 50 %). Fisk and Pidgeon's results support this prediction: fallacy rates for LU conjunctions were greater than 50 % for all but one LU conjunction in their data, with that conjunction having a fallacy rate of 48.4 % (Table 2) and with the average fallacy rate for LU conjunctions being over 50 % (Table 3).

For LL conjunctions, we assume $P(A) > P(B) > 0.5$ (both *A* and *B* are likely). Here again, the difference between $P(A \wedge B)$ and *P(B)* will tend to be small and will decline as *P(A)* rises (for the same reason given earlier). In this case, however, $P(A \wedge B) < 0.5$ does not hold, because if $P(A \wedge B)$ is close to *P(B)* and $P(B) > 0.5$, then $P(A \wedge B)$ 0.5. In this case, the model predicts the conjunction fallacy to be relatively frequent but always less than 50 % (because rates over 50 % can only occur when $P(A \wedge B) < 0.5$, which does not hold here). Fisk and Pidgeon's results again support this prediction: fallacy rates for LL conjunctions were less than 50 % for all but one LU conjunction in their data, with that conjunction having a fallacy rate of 51.1 % (Table 2); the average fallacy rate for LU conjunctions was less than 50 % (Table 3).

Finally, for UU conjunctions, we assume $P(B) < P(A) < 0.5$ (both *A* and *B* are unlikely). In this case, $P(A \wedge B)$ may be close to *P(B)* in some cases, but may be far from *P(B)* in other cases. For UU conjunctions, $P(A \wedge B) < 0.5$ necessarily holds, because $P(A \wedge B) < P(B) < 0.5$. Thus, the noise model predicts that high fallacy rates (rates over 50 %) may occur for UU conjunctions, but only in those cases where $P(A \wedge B)$ happens to be close to *P(B)*. For other UU conjunctions, fallacy rates will be lower (less than 50 %). Again, Fisk and Pidgeon's results support this prediction: fallacy rates for UU conjunctions were less than 50 % for all but one UU conjunction: the UU conjunction with a fallacy rate over 50 % (the UU conjunction in scenario 5 in Table 2) is also the UU conjunction with the smallest difference between conjunction and constituent probability estimates.

The probability theory plus noise model thus predicts that fallacy rates will be highest for LU conjunctions and will be lower for LL and UU conjunctions. The model does not make a direct prediction about the relative fallacy rates for LL and UU conjunctions: these fallacy rates will depend on the actual conjunctive and constituent probabilities involved. We therefore tested the model's agreement with Fisk and Pigeon's results on fallacy rates for LU, LL and UU conjunctions by considering the rate at which the simulation produced conjunction fallacy responses for these different types of conjunction. The original simulation did, indeed, produce fallacy rates for these types of conjunction that followed the pattern seen by Fisk and Pidgeon: fallacy rates were highest for LU conjunctions, next highest for LL conjunctions and marginally lowest for UU conjunctions (Table 3). In our re-run of the simulation with a range of different values for *d* and $\Delta d$, we found that this pattern held in all cases where *d* was between 0.01 and 0.1 and where $\Delta d$ was equal to 0.01 or 0.02 and held for no values of *d* or $\Delta d$ outside these ranges. This result demonstrates that the model can produce results matching those of Fisk and Pidgeon for a range of different parameter values. Again, the fact that the match to Fisk and Pidgeon's data depended primarily on the value of $\Delta d$ is in line with our proposal that the observed variation in fallacy rates is a consequence of increased noise in conjunctive probability estimates.

Table 3. Observed and simulated fallacy rates for LL, LU and UU conjunctions

| Conjunction type | Single fallacy |  | Double fallacy |  |
|---|---|---|---|---|
|  | (Observed) | (Simulation) | (Observed) | (Simulation) |
| LL | 36.5 | 32.3 | 19.3 | 3.1 |
| LU | 61.2 | 57.2 | 2.9 | 0.0 |
| UU | 29.7 | 30.5 | 4.9 | 1.7 |

## VARIATION AND VALUES OF THE ADDITION LAW IDENTITY

Recall that our model predicts that values for the addition law identity will be distributed around and close to zero in people's probability judgements and that our extended model (with increased noise for conjunctions and disjunctions) makes the further prediction that values for the addition law identity will be distributed around zero in a way that is directly proportional to both $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ and $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$.

We test these predictions using data from a previous experiment examining the addition law (Experiment 1 in Costello and Watts (2014)). In this experiment, we asked 83 participants to estimate the likelihood of a range of weather events A,*B* and conjunctions and disjunctions of those events. Half of the participants were asked these questions in a 'probability format' (asked 'What is the probability that the weather will be rainy on a randomly selected day in Ireland?'), and half were asked in a 'frequency format' (asked to 'Imagine a set of 100 different days, selected at random. On how many of those 100 days do you think the weather in Ireland would be rainy?'). We expect our model's predictions to hold for each question format; however, because the 'frequency format' questions are closer to the underlying counting mechanisms that our model assumes, we expect the match to predictions to be closer for this format. Similarly, because we assume a lower rate of random noise for single events than for conjunctions and disjunctions, we expect the single-event expression $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ to be less subject to random variation and so a more accurate predictor of variation in addition law values.

**DOI**: 10.1002/bdm

**Addition law results**

Table 4 shows the average values of $\langle P_E(A)\rangle, \langle P_E(B)\rangle$, $\langle P_E(A \wedge B)\rangle$ and $\langle P_E(A \vee B)\rangle$ for each pair of events in experiment for both frequency format and probability format groups, along with the average value for the addition law for each pair and the value of the expressions $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ and $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$ for that pair. Values of the addition law were distributed closely around 0 as predicted by the model in both groups, with overall means of $-0.007$ in both groups. We used one-sample $t$-tests across all individual values for the addition law expression in the frequency group ($t(491) = 0.537$) and in the probability group ($t(503) = 0.429$) to calculate Bayes factors in favour of or against the null hypothesis of a mean of 0 (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Both analyses gave strong evidence in favour of the null hypothesis (JZS Bayes factor =24.1 in the frequency group and JZS Bayes factor =25.7 in the probability group), as predicted by the model.

While values of the addition law were distributed closely around 0, these values did vary from one $A, B$ pair to another. In our model, this variation is expected to follow the terms $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ and $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$. There were reliable overall correlations between addition law values and the $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ and $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$ expressions ($r = 0.58$, $p < 0.01$ and $r = 0.43, p < 0.05$, respectively). As expected, correlations were higher in the frequency format group than in the probability format group and higher for the single-event expression $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ than for the $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$ expression.

Our model predicts that the sign of the addition law value should follow the sign of $P_E(A) + P_E(B) - 1$ (the addition law being positive when this expression is positive and negative when it is negative). The results supported this prediction. Of the 24 pairs of values for these expressions in Table 4, there were 10 cases where both were positive, 9 cases where both were negative and 5 where they differed in sign ($p < 0.05$ in Fisher's exact test). Our model makes the same prediction for $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$ and the addition law. This prediction was also supported: of the 24 pairs of values for these expressions, eight were both positive, nine were both negative and seven differed in sign ($p < 0.05$ in Fisher's exact test).

Table 4. Average probability estimates for pairs of weather events from participants in the 'frequency format' and 'probability format' groups in Experiment 1, Costello and Watts (2014) and average values of the addition law identity for each pair of events

| A | B | $P_E(A)$ | $P_E(B)$ | $P_E(A \wedge B)$ | $P_E(A \vee B)$ | addition law | $P_E(A) + P_E(B) - 1$ | $P_E(A \wedge B) + P_E(A \vee B) - 1$ |
|---|---|---|---|---|---|---|---|---|
| | | | | Frequency format group ($n = 41$) | | | | |
| Cloudy | Cold | 0.77 | 0.76 | 0.68 | 0.08 | 0.05 | 0.53 | 0.48 |
| Cold | Windy | 0.76 | 0.72 | 0.63 | 0.76 | 0.09 | 0.48 | 0.39 |
| Sunny | Cold | 0.39 | 0.76 | 0.42 | 0.82 | −0.09 | 0.15 | 0.24 |
| Cold | Thundery | 0.76 | 0.13 | 0.22 | 0.67 | 0.00 | −0.11 | −0.11 |
| Cloudy | Frosty | 0.77 | 0.31 | 0.33 | 0.07 | 0.04 | 0.07 | 0.04 |
| Frosty | Windy | 0.31 | 0.72 | 0.33 | 0.59 | 0.01 | 0.02 | −0.08 |
| Sunny | Frosty | 0.39 | 0.31 | 0.25 | 0.55 | −0.01 | −0.31 | −0.21 |
| Frosty | Thundery | 0.31 | 0.13 | 0.15 | 0.35 | −0.07 | −0.57 | −0.05 |
| Cloudy | Sleety | 0.77 | 0.02 | 0.33 | 0.66 | −0.03 | −0.04 | −0.01 |
| Sleety | Windy | 0.02 | 0.72 | 0.27 | 0.59 | 0.05 | −0.09 | −0.14 |
| Sunny | Sleety | 0.39 | 0.02 | 0.16 | 0.45 | −0.02 | −0.42 | −0.39 |
| Sleety | Thundery | 0.02 | 0.13 | 0.16 | 0.26 | −0.01 | −0.68 | −0.58 |
| Mean of addition law values | | | | | | −0.007 | | |
| Correlation with addition law values | | | | | | | 0.65* | 0.52 |
| | | | | Probability format group ($n = 42$) | | | | |
| Cloudy | Cold | 0.74 | 0.73 | 0.67 | 0.77 | 0.02 | 0.46 | 0.44 |
| Cold | Windy | 0.73 | 0.71 | 0.66 | 0.73 | 0.05 | 0.44 | 0.39 |
| Sunny | Cold | 0.39 | 0.73 | 0.47 | 0.75 | −0.01 | 0.12 | 0.22 |
| Cold | Thundery | 0.73 | 0.02 | 0.33 | 0.06 | 0.00 | −0.07 | −0.07 |
| Cloudy | Frosty | 0.74 | 0.32 | 0.37 | 0.64 | 0.05 | 0.06 | 0.01 |
| Frosty | Windy | 0.32 | 0.71 | 0.41 | 0.06 | 0.03 | 0.03 | 0.00 |
| Sunny | Frosty | 0.39 | 0.32 | 0.31 | 0.51 | −0.01 | −0.29 | −0.18 |
| Frosty | Thundery | 0.32 | 0.02 | 0.22 | 0.35 | −0.05 | −0.48 | −0.43 |
| Cloudy | Sleety | 0.74 | 0.28 | 0.04 | 0.62 | 0.00 | 0.02 | 0.02 |
| Sleety | Windy | 0.28 | 0.71 | 0.35 | 0.59 | 0.05 | 0.00 | −0.06 |
| Sunny | Sleety | 0.39 | 0.28 | 0.02 | 0.49 | −0.01 | −0.32 | −0.31 |
| Sleety | Thundery | 0.28 | 0.02 | 0.21 | 0.29 | −0.01 | −0.51 | −0.05 |
| Mean of addition law values | | | | | | −0.007 | | |
| Correlation with addition law values | | | | | | | 0.46 | 0.31 |
| Overall correlation with addition law values | | | | | | | 0.58** | 0.43* |

Average values for the expressions $\langle P_E(A)\rangle + \langle P_E(B)\rangle - 1$ and $\langle P_E(A \wedge B)\rangle + \langle P_E(A \vee B)\rangle - 1$ are also shown. Overall, both expressions had a significant positive correlation with addition law values, as predicted.

*$p < 0.05$.

**$p < 0.01$.

DOI: 10.1002/bdm

These results support our model's predictions about the addition law: values for that identity were close to zero for all individual conjunctions, and values varied around zero in a way that matches the predictions of the extended $d + \Delta d$ model. These results distinguish our model's account for the addition law identity from the weighted averaging account (Nilsson *et al.,* 2009; Nilsson *et al.,* 2016), which does not make these predictions.

### Comparing variance of single and conjunctive events

We also use the data from this experiment to compare the degree of variability in people's probability estimates for single and conjunctive events. Table 5 shows the SDs in people's estimates $P_E(A)$, $P_E(B)$ and $P_E(A \wedge B)$ for the 12 pairs of events in the experiment for both the frequency format and the probability format groups. There are two points to note about the data in this table. First, the degree of variability (SD) in people's estimates for events $A$, $B$ and $A \wedge B$ was reliably lower for estimates in the frequency format group than for estimates in the probability format group. Of the 36 possible comparisons (12 event pairs by three estimates: $P_E(A)$, $P_E(B)$ and $P_E(A \wedge B)$), 34 comparisons (95 %) showed lower variability (lower SD) in the frequency format group. This is broadly consistent with the approach taken in our model: because our model assumes that probability estimates are fundamentally frequentist (based on counting event occurrence), we would expect there to be less noise in frequency estimates, and so less variability in those estimates.

The second point to note is that the degree of variability (SD) in people's estimates for single events $A$ and $B$ tended to be lower than the degree of variability in estimates for conjunctions $A \wedge B$. This pattern was particularly evident in the frequency format group, where the SDs of both constituents $A$ and $B$ were less than the SD of the conjunction $A \wedge B$ in 10 out of 12 event pairs (the average SD value for single events in this group was 0.16 against an average SD value for conjunctions of 0.22). This pattern was less evident in the probability format group, where it only held for six out of 12 pairs (and where there was little difference in the averages). Because an initial assessment showed that probability estimates for events in this experiment were approximately normally distributed, we carried out an indicative statistical test to compare conjunctive and constituent variability using the $F$-test for equality of variance. In the frequency format group, the $F$-test for equality of variance ($df1 = df2 = 40$) showed a significant difference in variance in the predicted direction in 58 % of comparisons ($p < 0.05$), with the difference being somewhat close to significant (around $p = 0.10$) in another 17 % of comparisons. A similar comparison for the probability format group ($df1 = df2 = 41$), however, showed almost no significant differences ($p < 0.05$ in only 12.5 % of comparisons). This smaller difference in variability between single and conjunctive events in the probability format group may simply be a consequence of the higher overall degree of variability in that group. In general, these results are consistent with the idea that random variation is higher in probability estimates for conjunctions than those for single events (with this difference becoming more visible in cases were the overall level of variability is lower).

## DISCUSSION

Our aim in this paper has been to show that a detailed version of our probability theory plus noise model can explain the high rates of conjunction fallacy occurrence seen for some materials such as Tversky and Kahneman's Linda. This detailed version is, we think, both natural and reasonable: it reflects the fact that complex expressions like conjunctions and disjunctions may be more subject to random error, simply because they are more complex (and so provide more opportunity for such random error to occur). Our results show that high rates of occurrence of the conjunction fallacy cannot be taken as evidence that people's mechanisms for estimating probabilities do not follow probability theory: these high rates can arise purely because of random noise in such a

Table 5. Standard deviation in estimates for single and conjunctive weather events from participants in the 'frequency format' and 'probability format' groups in Experiment 1, Costello and Watts (2014) (mean values for these estimates are show in Table 4). SDs for single events were reliably lower than SDs for conjunctions, especially in the Frequency format group (where the SDs for $P_E(A)$ and for $P_E(B)$ were lower than the SD for $P_E(A \wedge B)$ in but 2 cases)

| | | Frequency format group ($n = 41$) | | | Probability format group ($n = 42$) | | |
|---|---|---|---|---|---|---|---|
| A | B | SD of $P_E(A)$ | SD of $P_E(B)$ | SD of $P_E(A \wedge B)$ | SD of $P_E(A)$ | SD of $P_E(B)$ | SD of $P_E(A \wedge B)$ |
| Cloudy | Cold | 0.12 | 0.13 | 0.19 | 0.23 | 0.19 | 0.23 |
| Cold | Windy | 0.13 | 0.17 | 0.21 | 0.19 | 0.17 | 0.24 |
| Sunny | Cold | 0.17 | 0.12 | 0.21 | 0.19 | 0.23 | 0.20 |
| Cold | Thundery | 0.12 | 0.12 | 0.24 | 0.23 | 0.21 | 0.25 |
| Cloudy | Frosty | 0.12 | 0.20 | 0.25 | 0.23 | 0.24 | 0.27 |
| Frosty | Windy | 0.20 | 0.17 | 0.23 | 0.24 | 0.17 | 0.26 |
| Sunny | Frosty | 0.17 | 0.20 | 0.22 | 0.19 | 0.24 | 0.24 |
| Frosty | Thundery | 0.19 | 0.12 | 0.19 | 0.24 | 0.21 | 0.21 |
| Cloudy | Sleety | 0.12 | 0.18 | 0.26 | 0.23 | 0.23 | 0.27 |
| Sleety | Windy | 0.18 | 0.17 | 0.25 | 0.23 | 0.17 | 0.25 |
| Sunny | Sleety | 0.17 | 0.18 | 0.19 | 0.19 | 0.23 | 0.19 |
| Sleety | Thundery | 0.18 | 0.12 | 0.17 | 0.23 | 0.21 | 0.16 |
| Average | | 0.16 | 0.16 | 0.22 | 0.22 | 0.21 | 0.23 |

SD, standard deviation.

rational reasoning process. Our results support this model in other ways by showing that this extended model can account for the relatively minor variation in values of the addition law around probability theory's required value of 0 and that variability in estimates for conjunctions tends to be higher than variability in estimates for single events.

### Scope of the model

What is the scope of our model? Because the model assumes that the probability *P(A)* is estimated by retrieving a random sample of episodes from memory and counting the number of *A*'s, it may seem that the model is only able to give probability estimates for events that have already been seen. This view 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 (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, an estimate of *P(A)* is produced by taking a random sample of episodes generated by constructive memory and counting the number 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 the similar agreement with probability theory for cancelling expressions such as the addition law in situations where we ask people to estimate probabilities for events that they have repeatedly experienced in the past (such as the weather events in our experiment), for past events that they have not directly experienced, for future events and for events that are to some degree imaginary. Experimental results support these predictions. For example, in an experiment asking participants to estimate the probability of people over the age of 60 years having certain diseases (such as Alzheimer's and diabetes), 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 gave a value for the addition law that was, on average, very close to 0 as required by probability theory. In an experiment asking participants to estimate the probability of a range of future events (such as a future increase in cigarette taxes and a future decline in smoking rates) and of various conjunctions and disjunctions of those events, we found that people's probability estimates gave values for the addition law that were, on average, very close to 0 as required by probability theory; similar results held for a number of other such 'cancelling' expressions (Costello & Watts, 2015b). 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 gave a value for the addition law that was, again, very close to 0 as required by probability theory. Together, these results suggest that our model applies to probability estimation for events in general and is not limited solely to events that have previously been seen.

### Factors influencing random error

The model we describe here makes three assumptions about random error in the proposed counting process behind probability estimation: that it is random, that it is symmetric and that it can be higher for complex expressions (such as conjunctions or disjunctions) than for the constituents of those expressions. We expect, however, that various general factors can influence this rate of random error in an experimental setting.

One such factor is simply individual differences between participants. We would expect that, when asking a group of participants to estimate the probability of some event *A*, some participants would have a higher rate of random noise (a higher value of *d*), while others would have a lower rate. Because in our model, the conjunction fallacy is a consequence of this random noise, we would expect higher conjunction fallacy rates for participants with higher values of *d* than for participants with lower values of *d*, all else being equal. Experimental results support this prediction. For example, Costello and Watts (2014) used various probabilistic expressions to estimate the value of *d* for individual participants in their experiments and found that this value was significantly correlated with the rate at which those participants committed the conjunction fallacy.

Another such factor is task demands. We would expect that more complex probability assessment tasks would involve higher rates of random error in probability assessment, while simpler tasks would have lower rates. Examples of more complex probability assessment tasks might involve those where stories or descriptions are provided to frame the required probability assessment and where participants must assess probabilities relative to those descriptions (as in Tversky and Kahneman's Linda example). Examples of less complex assessment tasks would be those that simply ask participants to assess probabilities directly, with no framing description, as our previous studies, which simply asked participants to estimate the probability of different types of weather such as 'rain', 'wind' and 'rain and wind' (Costello, 2009a; Costello & Watts, 2014). Because the rate of occurrence of the conjunction fallacy depends on the value of *d* in our model, we would expect higher fallacy rates for more complex tasks and lower rates for less complex tasks, all else being equal. We would also expect probability ranking tasks (where people are given a set of events and asked to simply choose the most probable) to be associated with higher rates of *d* than probability estimation tasks (where people are explicitly asked to estimate the probability of each event). This is because the requirement to produce an explicit probability estimate would, we think, cause people to focus more attention on the value of that estimate, so reducing the rate of

DOI: 10.1002/bdm

random error. Experimental results are approximately consistent with these expectations: for example, conjunction fallacy rates were lower in our weather tasks (which asked people to assess probabilities directly, with no framing description) than in Tversky and Kahneman's Linda task (which asked people to assess probabilities relative to a framing description about Linda's background and education). Fallacy rates are also typically lower in probability estimation tasks than in probability ranking tasks (Wedell & Moro, 2008).

Finally, a third possible factor affecting the rate of random noise is overall event complexity. Some probability assessment studies involve relatively simple events such as 'rain' or 'wind' (Costello, 2009a; Costello & Watts, 2014). Other studies, however, involve quite complex events; for example, Costello and Watts (2015b) asked people to assess the probability of events such as 'The Irish Government increases taxes on cigarettes in the next budget' and 'Smoking rates in Ireland decrease significantly in 2015'. We would expect such complex events to be associated with higher rates of random error $d$, simply because they are more complex and so have more 'points of failure': more points at which random error can have an impact. Again, experimental results are roughly consistent with this expectation: estimates for $d$ were noticeably higher for the complex events used in Costello & Watts, (2015a,c) than for the simpler events used in Costello (2009a) and Costello and Watts (2014).

Given this point about the relationship between overall event complexity and the degree of associated random noise, it is obvious that our model's assumption of one rate of error for single events and a higher rate for conjunctive events represents a fairly large simplification. In particular, some conjunctions involve two almost atomic constituents of relatively low complexity: for example the 'red triangle' and 'green square' conjunctions used by Zhao *et al.* (2009). Other conjunctions, such as 'The Irish Government increases taxes on cigarettes in the next budget *and* Smoking rates in Ireland decrease significantly in 2015', involve constituents that are themselves conjunctions of various other, simpler components. Examples such as these suggest that a sharp delineation between single events and conjunctions is not as straightforward as it initially appears and is assumed to be in our model.

It is clear from these suggestions about memory as simulation, about factors influencing noise and about the distinction between single events and conjunctions that our model is underspecified. Why do we not describe, in our model, exactly how memory works and precisely how these different levels of complexity and degrees of noise operate and interact? We could, for example, follow the approach taken in the Minerva-DM memory retrieval model of decision making (Dougherty *et al.,* 1999) or in Hilbert's 'noisy channel' account of biases in probability estimation (Hilbert, 2012). This underspecification in our model is intentional. Our aim in developing the probability theory plus noise account is to produce a model of probabilistic reasoning that is simple enough to produce clear and testable predictions about people's probability estimates. By deliberately describing the role of noise in an abstract, mathematical way, we are able to make such predictions. Both the Minerva-DM model and Hilbert's model are quite complex: each contains a range of different components, interacting in various different ways and controlled by various different parameters (see Costello & Watts, 2014, for comparisons between our model and these more complex alternatives). Such complex models are less amenable to the type of analysis needed to derive clear and testable predictions.

We see the development of the probability theory plus noise model as involving a series of increasing levels of approximation to the processes of human probability estimation, with each level of approximation introducing some important factor that was not included in previous levels. In this view, standard probability theory itself represents a zeroth-order approximation of human probabilistic reasoning: it includes (what we see as) the dominant factor in people's probability estimation (that is, the observed frequency of events) but does not include another important factor, which is noise in reasoning. The probability theory plus noise model represents a first-order approximation (it includes both the observed frequency of events and a basic approximation of noise in reasoning). The current model, then, represents a second-order approximation (it includes the observed frequency of events and a two-part approximation of noise in reasoning). Higher-order approximations would involve more detailed representations of noise including an account of the structure and complexity of the constituents involved in conjunctions. Our future work will involve further, incremental development within this framework.

## Theoretical position

Our theoretical proposal here is that human probabilistic reasoning is based on a fundamentally rational process (one that follows frequentist probability theory) that is subject to random noise. It is important to stress that we 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 judgements for some events. Instead, we propose that people's probability judgements 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.

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 our distinction between the rationality of a process (for probability estimation) and the

DOI: 10.1002/bdm

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 environment, 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 due to noise). 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 that 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 environment 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 terms $d$ and $\Delta d$ are both zero.

Our argument has broader implications for research on patterns of bias in aspects of people's probabilistic decision-making. A common pattern in such research is to identify a systematic bias in people's probability estimates and to then take that bias as evidence that people do not reason via the rules of probability theory but instead use some alternative, normatively incorrect, heuristic process. The conjunction fallacy is a major locus of this pattern: faced with the reliable occurrence of the conjunction fallacy in people's probability judgements, researchers have suggested that people estimate conjunctive probabilities using heuristics such as representativeness (Tversky & Kahneman, 1983), signed summation (Fisk & Pidgeon, 1996), configural weighted averaging (Nilsson *et al.,* 2009), pragmatic inference (Hertwig & Gigerenzer, 1999) and inductive confirmation (Crupi, Fitelson, & Tentori, 2008; Tentori *et al.,* 2013). Our results, however, suggest that this leap from an observed bias to an inferred heuristic (motivated by, and intended to explain, that bias) is premature. This is because random noise in reasoning can cause systematic biases in people's responses even when people are using normatively correct reasoning processes, and so there is little need to propose an alternative heuristic to explain those biases (see Budescu, Erev, & Wallsten, 1997; Erev *et al.,* 1994, for similar arguments). 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.

This position leads to a particular view on the motivation for alternative theories of probability estimation. It seems clear to us that the various alternative accounts (representativeness, or denominator neglect or other heuristic approaches) are motivated by the assumption that the observed biases and errors seen in people's probability judgements 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 do not 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 do not 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 follows the rules of probability theory, and our reasoning processes are necessarily subject to noise.

Our model is not the only way of accounting for biases in probability judgements while maintaining consistency with probability theory in the mechanism behind those judgements. The quantum probability model (Busemeyer & Bruza, 2012; Busemeyer, Pothos, Franco, & Trueblood, 2011) takes a somewhat similar approach. It describes a mechanism that starts with classical probability theory as its base but uses quantum mechanical ideas to add corrections that account for various observed biases (e.g. the conjunction fallacy) in certain situations. More precisely, the quantum probability expressions for $P(A)$ and $P(B)$ reduce exactly to standard probability theory when the events $A$ and $B$ are 'compatible', that is, when both probabilities can be estimated simultaneously using the same set of features. It is only when the events being estimated are 'incompatible' (that is, when they cannot both be estimated simultaneously using the same set of features) that this model deviates from standard probability theory by introducing quantum theoretic interference terms that, according to the model, produce the observed biases. While this model is in some ways quite close to our own (being based on standard probability theory, and reducing exactly to that theory in certain situations), it still suffers from the problem of motivation seen for heuristic accounts: as far as we can see, there is no a priori motivation for such quantum interference effects (beyond simply fitting the data on observed biases). This, again, is in contrast with our model, where the presence of noise in the model is motivated by the unavoidable fact of noise in both the world and the brain. To make a clear distinction between the quantum probability model and the noise model, we need to carry out empirical tests to assess the competing predictions of these accounts. We aim to do this in future work.

## CONCLUSIONS

The fundamental idea in our model is that people's process for estimating probabilities follows the requirements of probability theory and that the systematic biases away from probability theory seen in people's judgements are simply the consequence of random error in that process. In other

DOI: 10.1002/bdm

work we have shown that this model can explain biases such as conservatism, subadditivity and binary complementarity. We have also shown that for expressions in which this model predicts bias should be cancelled, people's probability estimates agree closely with the requirements of the probability theory just as predicted by the model (Costello & Watts, 2014). Here we have shown that this model can explain biases such as the conjunction fallacy and, in particular, can explain the high rates of conjunction fallacy occurrence for some conjunctions. Taken together, our results give evidence against the popular idea that people estimate probabilities using heuristics that do not follow the normative requirements of probability theory (Ariely, 2009; Gigerenzer & Gaissmaier, 2011; Kahneman, 2011; Shafir & Leboeuf, 2002).

## REFERENCES

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

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language*, *19*(2), 199–210.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. part i: New theoretical developments. *Journal of Behavioral Decision Making*, *10*(3), 157–171.

Busemeyer, J. R., & Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge University Press.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193.

Carlson, B., & Yates, J. (1989). Disjunction errors in qualitative likelihood judgment. *Organizational Behavior and Human Decision Processes*, *44*(3), 368–379.

Costello, F. (2009a). Fallacies in probability judgments for conjunctions and disjunctions of everyday events. *Journal of Behavioral Decision Making*, *22*(3), 235–251.

Costello, F. (2009b). How probability theory explains the conjunction fallacy. *Journal of Behavioral Decision Making*, *22*(3), 213–234.

Costello, F. J. & 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.

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

Costello, F. & Watts, P. (2015a). People's conditional probability judgments follow probability theory (plus noise). Submitted for publication.

Costello, F. & Watts, P. (2015b). Surprising rationality in people's probability estimation: Assessing two competing models of probability judgment. Submitted for publication.

Costello, F., & Watts, P. (2016). Probability theory plus noise: Replies to Crupi and Tentori (2015) and to Nilsson, Juslin and Winman (2015), in press. *Psychological Review*.

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

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

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

Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review*, *4*(1), 96–101.

Fisher, C. R., & 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., & 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.

Gavanski, I., & Roskos-Ewoldsen, D. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, *61*(2), 181.

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

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–306.

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

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856–874.

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

Nilsson, H., Juslin, P., & Winman, A. (2016). Heuristics can produce surprisingly rational probability estimates: A commentary on Costello and Watts (2014). *Psychological Review*, in press.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, *138*(4), 517.

Peterson, C., & Beach, L. (1967). Man as an intuitive statistician. *Psychonomic Bulletin*, *68*(1), 29–46.

Pidgeon, N., & Fisk, J. E. (1998). Conditional probabilities, potential surprise, and the conjunction fallacy. *The Quarterly Journal of Experimental Psychology: Section A*, *51*(3), 655–681.

Reyna, V. F. (2004). How people make decisions that involve risk a dual-processes approach. *Current Directions in Psychological Science*, *13*(2), 60–66.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

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

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

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory and Cognition*, *30*(2), 191–198.

Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., & Wen, J. (2003). The conjunction effect: New evidence for robustness. *American Journal of Psychology*, *116*(1), 15–34.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science*, *28*(3), 467–477.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*(1), 235.

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

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

**DOI**: 10.1002/bdm

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: effects of response mode, conceptual focus, and problem type. *Cognition*, *107*(1), 105–136.

Wolfe, C. R., & Reyna, V. F. (2010). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making*, *23*(2), 203–223.

Zegers, F. E. (1986). A family of chance-corrected association coefficients for metric scales. *Psychometrika*, *51*(4), 559–562.

Zegers, F. E., & Ten Berge, J. M. (1985). A family of association coefficients for metric scales. *Psychometrika*, *50*(1), 17–24.

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

**DOI**: 10.1002/bdm