Random variation and systematic biases in probability estimation

Rita Howe⁎, Fintan Costello

School of Computer Science, University College Dublin, Ireland

ARTICLE INFO

Keywords:
Variance
Noise
Probability estimation
Conjunction fallacy
Disjunction fallacy

ABSTRACT

A number of recent theories have suggested that the various systematic biases and fallacies seen in people's probabilistic reasoning may arise purely as a consequence of random variation in the reasoning process. The underlying argument, in these theories, is that random variation has systematic regressive effects, so producing the observed patterns of bias. These theories typically take this random variation as a given, and assume that the degree of random variation in probabilistic reasoning is sufficiently large to account for observed patterns of fallacy and bias; there has been very little research directly examining the character of random variation in people's probabilistic judgement. We describe 4 experiments investigating the degree, level, and characteristic properties of random variation in people's probability judgement. We show that the degree of variance is easily large enough to account for the occurrence of two central fallacies in probabilistic reasoning (the conjunction fallacy and the disjunction fallacy), and that level of variance is a reliable predictor of the occurrence of these fallacies. We also show that random variance in people's probabilistic judgement follows a particular mathematical model from frequentist probability theory: the binomial proportion distribution. This result supports a model in which people reason about probabilities in a way that follows frequentist probability theory but is subject to random variation or noise.

1. Introduction

Researchers over the last 50 years have identified a large number of systematic biases in people's judgments of probability. These biases are typically taken as evidence that people do not follow the normative rules of probability theory when estimating probabilities, but instead use a series of heuristics (mental shortcuts or 'rules of thumb') that sometimes yield reasonable judgments but sometimes lead to severe and systematic errors, causing the observed biases (Kahneman & Tversky, 1982). This 'heuristics and biases' view has had a major impact in psychology (Kahneman & Tversky, 1982; Gigerenzer & Gaissmaier, 2011), economics (Camerer, Loewenstein, & Rabin, 2003, 2003), law (Korobkin & Ulen, 2000; Sunstein, 2000), medicine (Dawson & Arkes, 1987; Eva & Norman, 2005) and other fields, and has influenced government policy in a number of countries (Oliver, 2013; Vallgårda, 2012).

Evidence for these systematic biases in people's probabilistic reasoning is very strong. The conclusion that these biases necessarily demonstrate heuristic reasoning processes is, however, less sure. Various researchers have shown that these biases may simply be a consequence of random variation or 'noise' in otherwise rational and normatively correct processes: random variation that produces systematic, directional effects (see e.g. Hilbert, 2012; Johnson, Blumstein, Fowler, & Haselton, 2013; Costello & Watts, 2014; Marchiori, Di Guida, & Erev, 2015; Costello & Watts, 2016). Support for this view comes from results showing that when people's individual, systematically biased, probabilistic judgements are combined in ways which statistically cancel out noise, those judgements tend to agree closely with the requirements of normative probability theory with no remaining systematic deviation (Costello,
While both the heuristic and the random variation approaches can explain observed patterns of bias in probabilistic reasoning, these accounts differ in their predictions about the consistency of such bias. The random variation approach necessarily predicts a large degree of inconsistency in responses such that if a person is biased on one presentation of a given item, they may not be biased on another. The heuristic or ‘rule of thumb’ account typically does not consider internal variation in responses or make provision for changes in response to the same stimuli. Representativeness accounts of heuristics, for instance, can account for ‘external’ variance - that is, fallacy responses will vary between different problems as representativeness covaries with frequency (Kahneman & Tversky, 1982). However, it makes no such argument for responses to the same problem. Early in heuristics research, Kahneman and Tversky (1982) rejected the notion of an approach that included responses perturbed by error.

Indeed, the evidence does not seem to support a “truth plus error” model, which assumes a coherent system of beliefs that is perturbed by various sources of distortion and error. Hence we do not share Dennis Lindley’s optimistic opinion that “inside every incoherent person there is a coherent one trying to get out,” and we suspect that incoherence is more than skin deep (Kahneman & Tversky, 1982, p. 313).

More recent approaches to heuristics argue that a “toolbox” of strategies may be used to solve problems under uncertainty (e.g. Rieskamp & Otto, 2006; Scheibehenne, Rieskamp, & Wagenmakers, 2013). This approach can produce variable responding but there is no consensus about how strategies are selected and evidence suggests that single-process models may be preferred over multiple-strategy models (Söllner, Bröder, Glöckner, & Betsch, 2014). To date there has been little research on the degree of variability in people’s probabilistic judgement: ‘noisy rational’ models of probabilistic reasoning simply assume random variation in people’s probability judgement, without investigating its extent or character. In this paper we aim to fill this gap in two ways. First, we give a mathematical model of the form and structure of variance in people’s probabilistic judgement; second we describe four experiments investigating the existence, characteristics, and properties of random variation in people’s probabilistic judgement, and on the relationship between this variance and systematic judgement bias. These experiments all focus on the occurrence of two particular systematic biases – the conjunction and disjunction fallacy – in simple tasks where people are asked to estimate the probability of constituent, conjunctive and disjunctive events in a presented set of events. These studies examine the degree of random variation in people’s estimates for these probabilities, the extent to which this random variation predicts conjunction and disjunction fallacy occurrence, and the degree to which fallacy responses are themselves randomly variable. These studies also examine specific theoretical predictions about the form which random variation will take in these tasks.

1.1. Biases in reasoning: the conjunction and disjunction fallacies

Perhaps the best-known and most studied bias in probabilistic reasoning is the conjunction fallacy, exemplified by the “Linda problem” of Tversky and Kahneman (1983). In this problem participants read the following statement about Linda:

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.

and then answer the following question:

Which is more probable?.
A. Linda is a bank teller.
A ∩ B. Linda is a bank teller and is active in the feminist movement.

Tversky and Kahneman (1983) found that over 80% of their participants judged A ∩ B as more likely than A in this and many similar problems. This response violates probability theory, which requires that P(A ∩ B) ≤ P(A) and P(A ∩ B) ≤ P(B) must always hold, simply because A ∩ B cannot occur without A or B themselves occurring. The conjunction, A ∩ B, under the probabilistic laws, cannot be more likely than the single constituent A, thus when a participant chooses the conjunction A ∩ B as more probable, they are committing a fundamental violation of rational probabilistic reasoning referred to as the ‘conjunction fallacy’.

A similarly reliable disjunction fallacy occurs when participants judge the constituents A,B as more likely than the disjunction, A ∨ B (Carlson & Yates, 1989; Bar-Hillel & Neter, 1993). These widely replicated fallacy results were taken as an indication that humans do not reason in a normative fashion; that is, they don’t apply probabilistic rules to real-life contexts. Instead, it was suggested that people employ heuristics or mental short cuts to solve these problems. The conjunction fallacy, for instance, was suggested to occur because people employed a “representativeness heuristic” when reasoning about conjunctive problems (Tversky & Kahneman, 1983). Under this theory, the fallacy occurs as the person described in the conjunction, A ∩ B, is more representative of the information presented in the character sketch than the person described by the constituent, A. However, a number of studies has called the validity of the heuristics account into question (Bonini, Tentori, & Osherson, 2004; Sides, Osherson, Bonini, & Viale, 2002). Experiments that manipulated class inclusion, for instance, demonstrated that the fallacy occurs regardless of whether the conjunction is representative or not (Gavanski & Roskos-Ewoldsen, 1991). Other studies have varied response mode - force choice vs estimation - or conceptual focus - frequencies vs probabilities - and found that it can greatly effect the fallacy rates observed (Wedell & Moro, 2008; Tversky & Kahneman, 1983; Hertwig & Gigerenzer, 1999; Fiedler, 1988; Reeves & Lockhart, 1993). More importantly, by manipulating probability values, fallacy rates of 10% to 85% can be found. Fisk and Pidgeon (1996) demonstrated that very high fallacy rates occur where P(A) was high and P(B) was low and very low fallacy rates will occur where both P(A) and P(B) were low. While fallacy rates are generally quite high, a frequent observation among this research is that a small number of participants do not seem overly susceptible to the fallacy. Over a number of conjunction problems, participants rarely have 100% error rates (Stolarz-Fantino, Fantino, Zizzo, & Wen, 2003).
1.2. Variability and cognitive biases

A number of formal probabilistic models have sought to show that a range of biases can be explained as a function of quasi-rational probabilistic reasoning instead of a heuristic process. These models have emphasized the role of random variation, or noise, in the decision-making process. Erev, Wallsten, and Budescu (1994) proposed a model to explain the observation that underconfidence (conservatism) and overconfidence could often be observed in the same judgement tasks. They demonstrated that subjective probability estimates perturbed by error can give this pattern of under- and overconfidence, even when judgements are accurate (also see Budescu, Erev, & Wallsten, 1997). Similarly, Hilbert (2012) proposed a theoretical framework based on noisy information processing. Under this framework, memory-based processes convert observations stored in memory into decisions. By assuming that these processes are subject to noisy variation and that this variation generates systematic patterns of error in decision-making, this approach explains a number of cognitive biases.

These models, however, simply assume the existence of random variation or noise in probabilistic reasoning; they do not describe the form and structure of this variation. Our main theoretical contribution in this paper is to give a mathematical description of variance in probabilistic reasoning. We take as our starting point a general model of noise in a normatively correct reasoning process: the probability theory plus noise model (PTN). This model assumes that people estimate probabilities via a mechanism that is fundamentally rational (following standard frequentist probability theory), but is perturbed in various ways by the systematic effects or biases caused by purely random noise or error. This approach follows a line of research leading back at least to Thurstone (1927) and continued by various more recent researchers (see, e.g. Bearden & Wallsten, 2004; Dougherty, Gettys, & Ogden, 1999; Erev, Wallsten, & Budescu, 1994; Hilbert, 2012). This model explains a wide range of results on bias in people’s direct and conditional probability judgments across a range of event types, and identifies various probabilistic expressions in which this bias is ‘cancelled out’ and for which people’s probability judgments agree with the requirements of standard probability theory (see Costello & Mathison, 2014; Costello & Watts, 2014, 2016, 2017, 2018, 2019; Costello, Watts, & Fisher, 2018).

In standard frequentist 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 to give a sample proportion. The expected value of these estimates is \( P(A) \), the probability of \( A \); individual estimates will vary with a ‘binomial proportion’ distribution around this expected value (taking \( N \) to be the sample size, the binomial proportion distribution is simply equal to the binomial distribution \( \text{Bin}(N, P(A)) \), rescaled by \( 1/N \) to represent sample proportions; see below). The probability theory plus noise model assumes that people estimate the probability of some event \( A \) in exactly the same way: by randomly sampling items from memory, counting the number that are instances of \( A \), and dividing by the sample size. If this process was error-free, people’s estimates would be expected to have an average value of \( P(A) \). Human memory is subject to various forms of random error, however. To reflect this the model assumes that events have some chance \( d < 0.5 \) of randomly being read incorrectly: there is a chance \( d \) that a ‘\( \neg A \)’ (not \( A \)) event will be incorrectly counted as \( A \), and the same chance \( d \) that an \( A \) event will be incorrectly counted as ‘\( \neg A \)’. We take \( P_r(A) \) to represent the probability that a single randomly sampled item from this population will be read as an instance of \( A \) (subject to this random error in counting). Since a randomly sampled event will be counted as \( A \) if the event truly is \( A \) and is counted correctly (this occurs with a probability \( 1 - d \) \( P(A) \)), since \( P(A) \) events are truly \( A \) and events have a \( 1 - d \) chance of being counted correctly), or if the event is truly \( \neg A \) and is counted incorrectly as \( A \) (this occurs with a probability \( 1 - P(A) \)) \( d \), since \( 1 - P(A) \) events are truly \( \neg A \), and events have a \( d \) chance of being counted incorrectly), the population probability of a single randomly sampled item being read as \( A \) is

\[
(P_r(A)) = (1 - d)P(A) + (1 - P(A))d = (1 - 2d)P(A) + d
\]

(1)

This equation gives the expected value or predicted average for people’s estimates for the probability of some event \( A \). Since individual estimates are produced via sampling, individual probability estimates will vary randomly around this expected value in an approximately binomial proportion distribution. Note that this predicted average embodies a regression towards the center, due to random noise: estimates are systematically biased away from the ‘true’ probability \( P(A) \), such that on average estimates will tend to be greater than \( P(A) \) when \( P(A) < 0.5 \), and will tend to be less than \( P(A) \) when \( P(A) > 0.5 \), and will tend to equal \( P(A) \) when \( P(A) = 0.5 \). This expression represents the expected value or average of people’s probability estimates for some event \( A \). Since this model of probability estimation gives a central role to random noise (and sampling) it does not predict that all probability estimates will exactly equal the value given in this expression. Instead, the prediction is that, since individual estimates are produced via sampling and are subject to random error, individual estimates will vary randomly around this expected value.

Original versions of this account assumed the same rate of random errors for all events (Costello & Watts, 2014). More recent versions (Costello & Watts, 2018, 2016, 2018) proposed a higher rate of this random error in complex events (conjunctions \( A \land B \) and disjunctions \( A \lor B \)). This extension allowed for increased regression in complex events, and was primarily intended to explain the wide range of conjunction and disjunction fallacy rates observed in the literature (ranging from 0% fallacy rates for some conjunctions to over 70%) in some cases this increased regression would push conjunctive estimates \( P(A \land B) \) closer to 0.5 than constituent estimates \( P_r(A) \), producing high conjunction fallacy rates for that conjunction. With this extension the model gave a close fit to data on fallacy rates across the full observed range (Costello and Watts, 2017). This idea of increased error for conjunctive or disjunctive events 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. \( A \land B \)) that is a function of those two variables will have a higher rate of error than either variable on its own. To reflect this, the model assumes a rate of random error of \( d \) for single events but of \( d + \Delta d \) for conjunctions and disjunctions (where \( \Delta d \) represents a small increase in the rate of random error). The PTN then predicts that the expected value of a conjunction estimate will be:

\[
(P_r(A \land B)) = (1 - (2[d + \Delta d])P(A \land B) + [d + \Delta d]
\]

(2)
and that for a disjunction estimate will be:

\[
(P_e(A \lor B)) = (1 - (2d + \Delta d))P(A \lor B) + [d + \Delta d]
\] (3)

with individual estimates varying randomly around these expected values in a binomial proportion distribution.

These \( \Delta d \) expressions are simplifying approximations, and were simply taken as given in previous presentations of the PTN model. In the Appendix we extend this model by giving a specific model of the differential effects of random error on combined estimates \( P_e(A \land B) \) and \( P_e(A \lor B) \), and show that this more precise model can be well approximated by these \( \Delta d \) expressions. The more precise \( \Delta d \) expression assumes that counting for complex items can take place in two separate ways - some familiar complex items can be treated “intelligently” and counted as if they are simple events while other complex items will be treated “separately.” In separable cases, there are three possible sources of error: when counting items, when counting bins, and when counting \( A \land B \) or \( A \lor B \). This more specific model is quite complex: we use these simplifying \( \Delta d \) approximations in the main body of the paper for ease of presentation and to indicate that these error rates are themselves uncertain. Indeed, the main \( d \) term in this model is also a simplifying approximation, suggesting as it does the existence of a fixed rate of random error in probabilistic recall (in fact, we expect the error rate itself to vary randomly from moment to moment, depending on a range of extraneous factors).

1.3. Fallacy occurrence

The conjunction (and disjunction) fallacy arise in this model purely as a consequence of this random variation. Assuming without loss of generality that \( P(B) \leq P(A) \), the general idea is that a reasoner’s probability estimates for the probabilities of \( B \) and \( A \land B \) will both vary randomly around their expected values \( \langle P_e(B) \rangle \) and \( \langle P_e(A \land B) \rangle \). This random variation means that some individual estimates will occur where \( P_e(B) < P_e(A \land B) \), producing a conjunction fallacy response. The closer the expected values \( \langle P_e(B) \rangle \) and \( \langle P_e(A \land B) \rangle \) are to each other, the greater the chance of this fallacy response occurring. More specifically, this model predicts that the rate of conjunction fallacy responses will increase with the difference between average estimates

\[
\langle P_e(A \land B) \rangle - \langle P_e(B) \rangle = (1 - 2d)[P(A \land B) + [d + \Delta d] - (1 - 2d)P(B) - d]
\]

(being low when this difference is negative and high when it is positive). When this difference is negative we have \( \langle P_e(A \land B) \rangle < \langle P_e(B) \rangle \). Since individual estimates \( P_e(A \land B) \) and \( P_e(B) \) are both perturbed by random noise (which is equally likely to be positive or negative), when this difference is negative we expect that an individual estimate \( P_e(A \land B) \) will randomly fall above an estimate \( P_e(B) \) less than 50% of the time, producing a conjunction fallacy rate of less than 50%. Rearranging, we see that this difference will be positive when

\[
\Delta d[1 - 2P(A \land B)] > (1 - 2d)[P(B) - P(A \land B)]
\]

and when this inequality holds we expect that an individual estimate \( P_e(A \land B) \) will randomly fall above an estimate \( P_e(B) \) more than 50% of the time, producing fallacy rates of over 50% (and indeed as high as 85% or 90%) for some events. This model can thus account for the wide range of conjunction fallacy rates seen in experimental studies.

In a similar way the model predicts that the rate of disjunction fallacy responses will increase with the difference between average estimates

\[
\langle P_e(A) \rangle \land \langle P_e(A \lor B) \rangle = (1 - 2d)[P(A) + d - (1 - 2d + \Delta d)] - [d + \Delta d]
\]

(being low when this difference is negative and high when it is positive). Since \( P(A) - P(A \lor B) = P(A \land B) - P(B) \) we have

\[
\langle P_e(A) \rangle - \langle P_e(A \lor B) \rangle = (1 - 2d)[P(A \land B) - P(B)] - \Delta d[1 - 2P(A \lor B)]
\]

and we see that this model predicts that for a given pair of events \( A \) and \( B \), the rate of disjunction fallacy occurrence should be approximately equal to the rate of conjunction fallacy occurrence (subject to a small difference of order \( \Delta d \)).

1.4. The addition law

These conjunction and disjunction fallacy predictions both concern patterns of deviation from the requirements of normative probability theory. Interestingly, by combining these results we obtain a prediction of agreement with one particular requirement of normative probability theory: the addition law. The addition law states that

\[
P(A) + P(B) - P(A \land B) - P(A \lor B) = 0
\]

must hold for all events \( A \) and \( B \). If we just take a single noise rate of \( d \) across all forms of probability estimation, we get

\[
(P_e(A) + P_e(B)) - (P_e(A \land B)) - (P_e(A \lor B)) = (1 - 2d)[P(A) + P(B) - P(A \land B) - P(A \lor B)] + 2d - 2d = 0
\]

and the addition law identity should also hold in people’s probability estimates according to this model. Taking our more complex \( \Delta d \) expressions for conjunctions and disjunctions, we get

4
\[ (\langle P_e(A) \rangle + \langle P_e(B) \rangle - \langle P_e(A \land B) \rangle - \langle P_e(A \lor B) \rangle = 2\Delta d \left[ P(A \land B) + P(A \lor B) - 1 \right] \]
\[ = 2\Delta d \left[ P(A) + P(B) - 1 \right] \sim 0 \]

Since \(-1 \leq P(A) + P(B) - 1 \leq 1\) necessarily holds, this model predicts that the average or expected value for this identity in people’s judgements will fall within \(2\Delta d\) of zero and we expect the addition law to hold, on average, in people’s probability estimates just as it does in normative probability theory. Note that, as before, this Equation gives the expected value or predicted average the addition law when computed from people’s probability estimates for some pair of events \(A, B\). Since individual estimates are produced via sampling and are subject to random error, individual values for this identity are predicted to vary randomly around this expected value.

Note also that the terms in this addition law expression can be rewritten as
\[ \langle P_e(A \land B) \rangle - \langle P_e(A) \rangle = \langle P_e(B) \rangle - \langle P_e(A \lor B) \rangle \]
and so correspond exactly to the terms predicting conjunction and disjunction fallacy occurrence, in the previous section. This model thus predicts simultaneous patterns of deviation from and agreement with the normative requirements of probability theory (deviation in terms of conjunction and disjunction fallacy occurrence; agreement in terms of the addition law).

1.5. Variance of probability estimates

This PTN model simply assumes the existence of random variation in probabilistic reasoning. Here we extend this model to derive predictions about the characteristic properties and degree of variance that should hold in people’s probability estimates (if people are estimating probabilities via noisy sampling as that model proposes). As before, we assume that people estimate the probability of some event \(A\) by randomly sampling some set of items from memory, counting instances of \(A\) in the sample (subject to random error in counting), and dividing by sample size. The variance of the sample count \(X\) in this process can be modelled via the binomial distribution. In the binomial distribution the probability of getting \(x\) successes in a sample of size \(N\) with fixed probability of success \(p\) is given by

\[ P(x|N, p) = \binom{N}{x} p^x (1-p)^{N-x} \]

with the mean value of this sample count being

\[ \text{mean}(X) = \sum_{x=0}^{N} xP(x|N, p) = pN \]

and the variance of this sample count being

\[ \text{var}(X) = \sum_{x=0}^{N} (x - pN)^2 P(x|N, p) = Np(1-p) \]

(4)

Since the variance of any random variable is the average squared difference between values of that variable and its mean, the variance of the sample proportion \(p_e\) (that is, the variance of the proportion of successes in a sample) is given by

\[ \text{var}(p_e) = \sum_{x=0}^{N} \left( \frac{x}{N} - \frac{pN}{N} \right)^2 P(x|N, p) \]
\[ = \frac{1}{N^2} \sum_{x=0}^{N} (x - pN)^2 P(x|N, p) \]
\[ = \frac{Np(1-p)}{N} \]
\[ = \frac{p(1-p)}{N} \]

(5)

If people are estimating probabilities via the sampling process assumed in the probability theory plus noise model (where their probability estimate for some event \(A\) is equal to the proportion of items in a random sample that were counted as instances of \(A\), subject to random noise), then we would expect the variance of probability estimates to approximately follow this expression. More specifically, for event \(A\) and noise rate \(d\) we would expect the variance of people’s probability estimates \(P_e(A)\) to be

\[ \text{var}(p_e(A)) = \frac{\langle P_e(A) \rangle(1 - \langle P_e(A) \rangle)}{N} \]

(6)

where \(N\) is the sample size used when estimating probabilities and \(\langle P_e(A) \rangle = (1 - 2d)P(A) + d\) is the probability of an item being read as \(A\) (and where for conjunctive or disjunctive events we use \(d + \Delta d\), as before). Predicted Standard Deviation (SD) of probability estimates will then be the square root of this variance.

Given this theoretical background we now describe a series of experiments investigating the degree of random variation in probabilistic judgement, and the relationship between that variation and fallacy occurrence, in two different types of judgement task. Experiments 1 and 2 examine variance and fallacy occurrence in probability estimation for everyday events; Experiments 3 and 4 examine variance and fallacy occurrence in probability estimation for simple visual stimuli. The PTN predicts that the variance within
the probability estimates differs depending on whether the estimate is for a constituent, conjunction or disjunction, where higher variance should be observed for the complex statements. We will test this prediction and look at how variance in responses relates to fallacy rates; whether participants will produce fallacious responses repeatedly and whether they will be consistent or inconsistent in producing them across stimuli. Experiment 3 and 4 use stimuli with known objective probability values, allowing us to test predictions about the relationship between objective probability value and probability estimates, variance of estimates, and fallacy occurrence.

2. Experiment 1

Experiment 1 sought to investigate the variance in probability estimates using simple natural language estimation tasks. The participants were presented with single weather events (‘cold’, ‘rainy’) and conjunctive and disjunctive weather events (‘cold and rainy’, ‘cold or rainy’) and asked to estimate the probability or frequency of these weather events. The weather types were presented to participants in a randomised order and the participants were randomly assigned to one of two groups—frequency questions or probability questions.

This experiment will test a number of predictions about subjective estimates. The main impetus of this paper is to examine the variability in judgements. Here, we will investigate whether participants estimates agree with probability theory and whether those participants will produce noisier estimates for complex statements - conjunctions and disjunctions - than constituents. Theoretical approaches such as representativeness accounts and averaging models assume that participants do not produce estimates in line with probability theory while noise models such as the PTN do under certain circumstances. There is a theoretical divide here, broadly speaking, where the theories on extensional errors can be classified as those that propose that judgements are produced by a process radically different to probability theory and those that propose that judgements are produced by a process akin to probability theory. Representativeness accounts argue that participant judgements are not consistent with probability theory because they produce fallacies. The PTN, on the other hand, predicts that participants produce fallacies while being consistent with certain aspects of probability theory. We will investigate these claims.

An occasional finding in the literature is that question type (whether questions are about event frequency or event probability) affects the rate of fallacy production. We examine this factor here, and ask whether the question type with the higher fallacy rate will also have a higher degree of response variability.

2.1. Materials and method

The materials consisted of sets of questions about the likelihood (frequency, or probability) of a type of weather on a given day. Each set had 7 constituents, 8 conjunctions and 8 disjunctions (see Table 1 for materials). The questions were the same for each participant in each group but displayed in a randomised order. 94 participants were recruited from the student body in exchange for course credit, and were randomly assigned to either the frequency or the probability group. For the frequency group, the participants were asked *Imagine a set of 100 different days, selected at random. On how many of those 100 days do you think the weather in Ireland will be [weather type]?* Participants then indicated their answer using a scale of 0 to 100, where 0 indicated that they thought that there would be [weather type] on zero of those days, while 100 meant that they thought there would be [weather type] on 100 of those days. The probability group were asked *What is the probability that the weather will be [weather type] on a randomly selected day in Ireland?* Again, they indicated their answer on a scale of 0–100. Answers of 0 meant that the weather type would never happen, while answers of 100 meant that the weather type was certain to happen on a given day.

### Table 1

| A          | B              | A ∩ B | \( p_A \) | \( p_B \) | \( p_{A \cap B} \) | Total conj. fallacy count (/94) |
|------------|----------------|-------|----------|----------|------------------|-------------------------------|
| Warm       | Sunny          | Warm and Sunny | 0.32      | 0.33      | 0.26             | 32‡                           |
| Rainy      | Cold           | Rainy and Cold | 0.64      | 0.65      | 0.55             | 36†                           |
| Rainy      | Warm           | Rainy and Warm | 0.64      | 0.32      | 0.31             | 35‡                           |
| Sunny      | Windy          | Windy and Sunny | 0.33      | 0.62      | 0.33             | 43                            |
| Snowy      | Cloudy         | Snowy and Cloudy | 0.13      | 0.73      | 0.16             | 45                            |
| Cloudy     | Windy          | Windy and Cloudy | 0.73      | 0.62      | 0.61             | 43                            |
| Snowy      | Sunny          | Snowy and Sunny | 0.33      | 0.13      | 0.12             | 37‡                           |
| Cloudy     | Rainy          | Cloudy and Rainy | 0.73      | 0.64      | 0.59             | 36†                           |

Note: ‡ probability <0.01 in a cumulative binomial test with \( N = 94, p = 0.5 \).

† probability <0.05 in a cumulative binomial test with \( N = 94, p = 0.5 \).
2.2. Results

Under the PTN, violations of probability theory - conjunction and disjunction fallacies - should arise as a function of two things: probability values and variance. In the results below we examine how these variables contribute to fallacies. It is expected that the participant estimates should be consistent with elements of probability theory despite the production of fallacies. We will examine a number of things; whether judgements are consistent with the addition law and whether variability is greater for complex items than simple ones. Representativeness and noise accounts of the fallacies make disparate predictions about these items.

2.2.1. Response mode and fallacy rate

Previous experiments looking at response mode have typically found lower fallacy rates when participants are presented with conjunction and disjunction questions in a frequency format than a probability format. To test where a difference would exist, each conjunction in the frequency group was paired with the respective conjunction in the probability group and a 2-sample test for equality of proportions was calculated. This found no significant difference in fallacy rates for any of the pairs. The disjunctions in both groups were also paired in this fashion and again, the equality of proportions test was calculated. Again, there was no difference in the fallacy rates between the groups. As the two groups produced very similar estimates and fallacy rates, they were collapsed for the purpose of analysis.

2.2.2. Estimation and probability theory

We examined whether participant judgements could appear consistent with normative reasoning under certain conditions. An important prediction of the PTN is that participant judgements should be in line with the addition law even while producing fallacies. Under the PTN, the noise in participant judgements should cancel and produce a response that is in compliance with the addition law. The averaged values for each P(A), P(B), P(A\&B) and P(A\lor B) estimate were used to test this. The participants’ estimates showed good compliance with the addition law. The estimates were close to the expected mean value of 0, with mild deviations from this value. We found an overall value of 0.019 for the estimates. For the frequency group, the average estimate was 0.035. In the probability group, there was even closer compliance with the addition law. There, the average estimate was 0.006.

From the addition law, we observe that the sum of estimates for the positive terms, P(A), P(B), should equal the sum of estimates for the negative terms, P(A \& B), P(A \lor B). 

\[ P(A) + P(B) = P(A \& B) + P(A \lor B) \]

Using this, we constructed a scatterplot to investigate compliance for probability theory. Fig. 1 shows a scatterplot the positive and negative terms for both groups in experiment 1. A Deming regression was used to determine how consistent the individual estimates were with probability theory. If the participant estimates are consistent with probability theory, then this regression will produce a line of best fit that follows the line of identity. As the figure shows, values for the addition law are distributed approximately symmetrically around the line of identity, with the line of best agreeing closely with the line of identity, as predicted by our model. A JZS Bayes Factor analysis based on a paired t-test of x and y values in this scatterplot (positive terms and negative terms in the addition law) gave strong evidence in favour of the null hypothesis that x and y values were equal (Scaled JZS Bayes Factor = 24.5), supporting the conclusion that the addition law identity holds in individual participant probability estimates. This replicates a range of previous results on the addition law (Costello & Watts, 2014; Costello & Watts, 2016; Costello & Watts, 2018).

2.2.3. Addition law and fallacy rates

The conjunction fallacy rate relative to A should follow the disjunction fallacy rate relative to B, for any pairing of A,B. This arises as a natural consequence of the addition law, which the PTN predicts will be related to the fallacy rates. By rearranging the terms of the addition law, we see that

\[ P(A \& B) - P(A) = P(B) - P(A \lor B) \]

If the participant judgements are consistent with this prediction, then we should see analogous responses to conjunctions and disjunctions; for example, when the conjunction fallacy rate versus A is low, then the disjunction rate versus B should be low. In Table 2, we observe that the related fallacy rates P(A \& B) - P(A) and P(B) - P(A \lor B) are typically close. A very strong positive correlation was observed between the relative conjunction and disjunction fallacy rates, \( r = 0.912, \ p < 0.00001 \).

2.2.4. Variability and probability estimates

From the PTN, we expect that the average estimate difference between the constituent and complex item can be used to predict the resultant fallacy rate for that complex item. To test this prediction, \( \bar{P}_\ell(A \& B) - \bar{P}_\ell(A) \) was calculated for each constituent and conjunction and \( \bar{P}_\ell(B) - \bar{P}_\ell(A \lor B) \) was calculated for each constituent and disjunction. This was then compared to the fallacy rates. These values are shown in Table 2. A Pearson’s correlation was used to examine the relationship between estimate difference and total fallacy rate for each pairing. A strong positive correlation was observed the average conjunction fallacy rates and average calculated estimate difference, \( r = 0.77, \ p < 0.0005 \). For the disjunctions, a very strong positive correlation was observed between disjunction fallacy rates and estimate difference, \( r = 0.92, \ p < 0.00001 \).

\[ \text{Only one pair were significantly different from each other - cloudyvrainy, which had a fallacy rate of 50\% in the frequency group and 25\% in the probability group, respectively.} \]
To test whether these correlations held across different sets of participants, we performed 100 random split-half correlations, dividing participants into two randomly chosen equal-sized halves, calculating conjunctive and disjunctive fallacy rates for each pair of events $A, B$ in one half and calculating estimate differences for those pairs in the other half, and measuring the correlation between those measures. There was strong positive relationship between average estimate difference and conjunction fallacy rate (average $r = 0.66, \min r = 0.51, p < 0.0001$ in all cases) and between average estimate difference and disjunction fallacy rate, (average $r = 0.80, \min r = 0.65, p < 0.00001$ in all cases).

However, the average difference and the fallacy rate are two measures that are by definition connected - one is the measure of the number of times that $P(A \land B)$ exceeds $P(A)$ for the conjunction, the other a measure of, on average, how much $P(A \land B)$ is larger than $P(A)$.

**Table 2**

| Conjunction | Average Difference | Fallacy rate | Conjunction | Average Difference | Fallacy rate |
|-------------|-------------------|--------------|-------------|-------------------|--------------|
| Cloudy v Sunny | 0.59 | 4% | Snowy v Sunny | 0.56 | 5% |
| Rainy v Sunny | 0.36 | 6% | Warm v Sunny | 0.38 | 6% |
| Sunny v Snowy | 0.26 | 10% | Snowy v Sunny | 0.27 | 9% |
| Snowy v Sunny | 0.07 | 16% | Sunny v Snowy | 0.34 | 9% |
| Cloudy v Sunny | 0.16 | 22% | Snowy v Sunny | 0.19 | 22% |
| Cloudy v Warm | -0.21 | 21% | Snowy v Warm | 0.14 | 24% |
| Sunny v Cold | -0.12 | 21% | Sunny v Cold | 0.14 | 35% |
| Rainy v Cold | -0.15 | 22% | Snowy v Sunny | 0.15 | 35% |
| Cold | -0.18 | 27% | Sunny v Snowy | -0.11 | 36% |
| Cloudy | -0.12 | 28% | Snowy v Sunny | -0.13 | 37% |
| Warm v Sunny | -0.12 | 26% | Warm v Sunny | -0.13 | 38% |
| Sunny v Warm | -0.17 | 36% | Warm v Sunny | -0.15 | 39% |
| Cloudy v Sunny | -0.13 | 44% | Cloudy v Sunny | -0.12 | 44% |
| Winter v Sunny | -0.14 | 44% | Cloudy v Sunny | -0.10 | 55% |

To test whether these correlations held across different sets of participants, we performed 100 random split-half correlations, dividing participants into two randomly chosen equal-sized halves, calculating conjunctive and disjunctive fallacy rates for each pair of events $A, B$ in one half and calculating estimate differences for those pairs in the other half, and measuring the correlation between those measures. There was strong positive relationship between average estimate difference and conjunction fallacy rate (average $r = 0.66, \min r = 0.51, p < 0.0001$ in all cases) and between average estimate difference and disjunction fallacy rate, (average $r = 0.80, \min r = 0.65, p < 0.00001$ in all cases).

However, the average difference and the fallacy rate are two measures that are by definition connected - one is the measure of the number of times that $P(A \land B)$ exceeds $P(A)$ for the conjunction, the other a measure of, on average, how much $P(A \land B)$ is larger than $P(A)$.
than $P(A)$ for the conjunction. This also holds for the disjunction where the average difference is the measure of how much $P(A \lor B)$ is smaller than $P(A)$, while the fallacy rate is the measure of how many times $P(A \lor B)$ is less than $P(A)$. To address this, these measures were separated and used to predict fallacy rates.

The estimates from any participant that had produced a fallacy for a particular conjunction or disjunction was excluded and the average $P_\text{e}(A \land B) - P_\text{e}(A)$ or $P_\text{e}(B) - P_\text{e}(A \lor B)$ difference was calculated for the participants that had not produced the fallacy. For instance, if participants 3, 5, and 7 had produced a fallacy response for Cloudy vs Cloudy\&Snowy, their estimates were removed and the $P_\text{e}(A \land B) - P_\text{e}(A)$ value for Cloudy vs Cloudy\&Snowy was then calculated for the participants that had produced no fallacy. This procedure was then repeated for each conjunction and disjunction. Table 2 which displays the average difference calculated for the restricted set of estimates and how it relates to the conjunction and disjunction fallacy rate for that pair. Higher fallacy rates were observed when the differences were close to zero, while lower fallacy rates were observed where the differences were much lower than zero. Pearson’s correlations were again calculated for estimate difference and fallacy rate. A very strong positive correlation was found for the restricted estimate difference and the conjunction fallacy rate, $r = 0.96$, $p < 0.00001$. A strong positive relationship was also found for the disjunction rate and the restricted estimates, $r = 0.78$, $p < 0.0005$.

In general, greater variance was observed for the complex statements than the constituents. The conjunctions were more variable than their constituent counterparts on 81% of the occasions while the disjunctions were more variable on 56% of the occasions. In the probability group, 93% of the conjunctions were more variable than their constituents, while in the frequency group 69% of the conjunctions were more variable. For the disjunctions, the opposite pattern was observed, with 75% of the frequency group showing higher variance, while 38% of the probability group’s disjunctions were more variable. Levene’s test of homogeneity of variances\(^2\) was used to determine if any of these were more variable at statistically significant levels. In the conjunctions, 13% of them were significantly more variable at 0.05 level, while a further 13% for were significant at the 0.1 level. For the disjunctions, Levene’s test found that 13% were significantly variable at the 0.05 while a further 10% were variable at the 0.1 level.\(^3\)

To examine the relationship between probability estimates and variability in producing the fallacies, 95% confidence intervals were constructed for the constituent and complex item from the restricted estimates. Each instance of a fallacy response (where $P(A \land B) > P(B)$ or $P(A \lor B) < P(A)$ was observed) was removed and the confidence intervals were then constructed using the instances where no fallacy occurred. In Tables 3 and 4 these values are displayed, in addition to degree to which the two confidence intervals overlapped. These results demonstrate that for high fallacy rates to occur there must be an overlap in the confidence interval of the constituent and the complex statement. That is, that the constituent and conjunction or disjunction estimate must be close to each other. The closer the estimates get to each other, the more likely the fallacy is to result. Large negative overlaps result in very low fallacy rates while overlaps around or above 0 will result in fallacy rates of approximately 50%. The larger the positive overlap, the greater the fallacy rate will be. A strong positive correlation was observed between fallacy rate and confidence interval overlap, $r = 0.778$, $p < 0.00001$.

### 2.3. Testing averaging models of the conjunction fallacy

Finally, it is worth noting that results from this experiment pose a challenge for one type of heuristic-based approach to conjunctive probability estimation and the conjunction fallacy: an approach where conjunctive probability estimates are produced by averaging constituent probabilities. Approaches following this line initially proposed that the conjunction estimate was simply the mean of the two constituent probabilities (Carlson & Yates, 1989; Fantino, Kulik, Stolarz-Fantino, & Wright, 1997). More recently Nilsson and colleagues (Nilsson, Winman, Juslin, & Hansson, 2009) have proposed a more sophisticated configural cue model, where conjunctive probabilities are computed by a weighted average of the form

\[
P(A \land B) = W \min(P(A), P(B)) + (1 - W) \max(P(A), P(B)) \quad 0.5 \leq W \leq 1
\]  

where a higher weight is given to the lower constituent probability and a lower weight to the higher constituent. Disjunctive probabilities are computed by an analogous weighted average

\[
P(A \lor B) = (1 - W) \min(P(A), P(B)) + W \max(P(A), P(B)) \quad 0.5 \leq W \leq 1
\]

but with the assignments of weights reversed, so that a lower weight is given to the lower constituent probability and a higher weight to the higher constituent.

Note that these conjunctive and disjunctive probability values will satisfy the addition law and similar identities, and so this model is consistent with those results (Nilsson, Juslin, & Winman, 2014). Even with this configural weighting, however, the average of two numbers is always greater than the minimum of those two numbers and less than the maximum (except when the numbers are equal). This means that these averaging accounts predict that the conjunction probability will almost always be greater than the lower constituent probability: that the conjunction fallacy will occur for almost every conjunction (and that the disjunction probability will almost always be less than the higher constituent probability: that the disjunction fallacy will occur for almost every disjunction). This is clearly not the case: there are many conjunctions for which the fallacy does not occur at anything close to 100%. To address this problem,
Nilsson et al.'s model also includes a noise component that randomly perturbs conjunctive probability estimates, sometimes moving the conjunctive probability below the lower constituent probability and so eliminating the conjunction fallacy for that estimate (and similarly for disjunctions). Since this noise is random, it has at most a 50% chance of moving a conjunctive probability (produced by averaging) below its lower component probability. This 50% chance arises when constituent and conjunctive probabilities are equal: in all other cases the conjunctive probability is greater than its lower constituent and so the chance of the conjunctive estimate falling below the constituent probability is necessarily less than 50%. This means that this noisy averaging model necessarily predicts conjunction fallacy will be predominant (occurring at rates of 50% or higher) for all conjunctions (see Nilsson et al., 2009, p. 521).

We can carry out a conservative assessment of this prediction by using the cumulative binomial test to ask whether the total number of conjunction fallacy occurrences observed in our experiment is consistent with the hypothesis that conjunction fallacy responses occur with a probability of 0.5 (the minimum probability predicted in this ‘noisy average’ account). Applying the cumulative binomial test to the total conjunction fallacy counts given in Table 1 (with N = 94, since there were 94 participants in total, and p = 0.5) we find that the total fallacy rates for 5 out of 8 conjunctions are inconsistent with the noisy averaging hypothesis at the 0.05

### Table 3
Confidence Intervals for Conjunctions (Exp 1). The table below displays the confidence intervals for the conjunctions in experiment 1. 95% confidence intervals were constructed using the restricted estimates for the constituent and conjunctions. From this, we could calculate how much the probability estimates overlapped for each pair. A positive value shows that the estimates for the constituent and conjunction typically overlapped and had higher fallacy rates. A negative value meant that the estimates typically did not overlap and were associated with low fallacy rates. A strong positive correlation was observed between fallacy rate and confidence interval overlap, $r = 0.778, p < 0.00001$.

| P(A) | P(A ∨ B) 95% CI | P(A) 95% CI | P(A ∧ B) 95% CI | Overlap | Fallacy |
|------|----------------|-------------|----------------|---------|---------|
| Cloudy | Snowy Cloudy | 0.69 | 0.76 | 0.09 | 0.18 | −0.51 | 4% |
| Rainy | Rainy Warm | 0.60 | 0.69 | 0.24 | 0.33 | −0.27 | 6% |
| Sunny | Snowy Sunny | 0.30 | 0.37 | 0.04 | 0.10 | −0.20 | 10% |
| Windy | Windy Sunny | 0.60 | 0.68 | 0.25 | 0.33 | −0.27 | 11% |
| Cloudy | Windy Cloudy | 0.71 | 0.79 | 0.53 | 0.62 | −0.09 | 15% |
| Snowy | Snowy Sunny | 0.30 | 0.37 | 0.04 | 0.10 | −0.20 | 16% |
| Cloudy | Cloudy Rainy | 0.71 | 0.79 | 0.49 | 0.58 | −0.13 | 21% |
| Sunny | Warm Sunny | 0.30 | 0.38 | 0.19 | 0.25 | −0.05 | 21% |
| Rainy | Rainy Cold | 0.61 | 0.70 | 0.46 | 0.55 | −0.06 | 22% |
| Warm | Warm Sunny | 0.29 | 0.37 | 0.17 | 0.24 | −0.05 | 26% |
| Cold | Rainy Cold | 0.66 | 0.74 | 0.46 | 0.56 | −0.10 | 27% |
| Rainy | Cloudy Rainy | 0.61 | 0.71 | 0.50 | 0.59 | −0.02 | 28% |
| Warm | Rainy Warm | 0.31 | 0.41 | 0.15 | 0.24 | −0.07 | 36% |
| Windy | Windy Cloudy | 0.64 | 0.74 | 0.50 | 0.60 | −0.04 | 42% |
| Sunny | Windy Sunny | 0.32 | 0.43 | 0.19 | 0.27 | −0.05 | 44% |
| Snowy | Snowy Cloudy | 0.07 | 0.21 | 0.03 | 0.10 | 0.03 | 48% |

### Table 4
Confidence Intervals for Disjunctions (Exp 1). The table displays the confidence intervals for the frequency and probability groups in experiment 1. 95% confidence intervals were constructed for the restricted constituent and disjunction estimates. A positive overlap shows the degree that the estimates for the constituent and conjunction typically overlapped. Higher fallacy rates with a larger overlap. A negative value meant that the estimates typically did not overlap and were associated with low fallacy rates. A very strong positive correlation is observed between the CI overlap and fallacy rate, $r = 0.874, p < 0.00001$.

| P(B) | P(A ∨ B) 95% CI | P(A) 95% CI | P(A ∧ B) 95% CI | Overlap | Fallacy |
|------|----------------|-------------|----------------|---------|---------|
| Snowy | Snowy Cloudy | 0.07 | 0.13 | 0.62 | 0.71 | −0.49 | 5% |
| Warm | Rainy Warm | 0.26 | 0.33 | 0.64 | 0.71 | −0.31 | 6% |
| Sunny | Windy Sunny | 0.28 | 0.34 | 0.68 | 0.78 | −0.26 | 9% |
| Snowy | Snowy Sunny | 0.06 | 0.12 | 0.31 | 0.40 | −0.19 | 9% |
| Windy | Windy Cloudy | 0.53 | 0.62 | 0.73 | 0.80 | −0.11 | 22% |
| Rainy | Cloudy Rainy | 0.55 | 0.65 | 0.75 | 0.82 | −0.10 | 24% |
| Warm | Warm Sunny | 0.23 | 0.31 | 0.36 | 0.46 | −0.05 | 32% |
| Rainy | Rainy Cold | 0.52 | 0.62 | 0.67 | 0.76 | −0.05 | 34% |
| Cold | Rainy Cold | 0.54 | 0.65 | 0.71 | 0.78 | −0.06 | 35% |
| Sunny | Snowy Sunny | 0.26 | 0.35 | 0.37 | 0.47 | −0.02 | 36% |
| Cloudy | Cloudy Rainy | 0.63 | 0.72 | 0.77 | 0.84 | −0.05 | 37% |
| Sunny | Warm Sunny | 0.24 | 0.32 | 0.35 | 0.46 | −0.03 | 38% |
| Rainy | Rainy Warm | 0.51 | 0.62 | 0.66 | 0.76 | −0.04 | 39% |
| Cloudy | Windy Cloudy | 0.61 | 0.71 | 0.74 | 0.82 | −0.03 | 44% |
| Windy | Windy Sunny | 0.49 | 0.59 | 0.63 | 0.73 | −0.04 | 44% |
| Cloudy | Snowy Cloudy | 0.63 | 0.74 | 0.74 | 0.82 | 0.00 | 55% |
significance level, and 3 out of 8 are inconsistent at the $p = 0.01$ level. For the conjunction ‘Rainy and Cold’, for example, 36 out of 94 participants gave a conjunction fallacy response. Under the assumption that $P(\text{fallacy}) = 0.5$, the probability of observing a fallacy count of 36 or less in a sample of 94 responses is less than $p = 0.05$. Similar results hold for the disjunction fallacy.

2.4. Experiment 1 discussion

As predicted by the PTN, the participants’ estimates were consistent with probability theory (in terms of the addition law) while simultaneously deviating from probability theory (in terms of frequent occurrence of the conjunction and disjunction fallacies). Rates of occurrence of the conjunction and disjunction fallacy were closely connected to difference in estimates, variance and overlap measures (just as predicted by that model). Results showed that participants are typically more variable across a range of conjunction and disjunction estimates than they are for constituents while the range of fallacy rates observed for the pairings are in line with those observed in previous research (4–48% for conjunctions and 5–55% for disjunctions). This has important implications for the production of fallacies as the high fallacy rates seem to arise due to a combination of comparably high variance in the conjunction or disjunction and probability estimates for the constituent that consistently overlap with the conjunction or disjunction probability estimates. Lower fallacy rates are typically observed where the estimates are far apart.

Unlike other findings in the literature, there was no significant observed difference between the frequency and probability groups in their estimates or fallacy rates. This was a consistent observation for both the conjunctions and disjunctions, with most pairs only differing by a few percentage points between the two response modes and no statistically significant difference found upon analysis. Previous research had suggested that fallacy rates could be manipulated by varying the response mode. However, the stimuli set used was more complex than the simple events used here, which may account for the observed difference.

Experiment 1 has established three important things: participants can be consistent with probability theory and still commit the fallacies, that variance exists between question types for participants, and that participants can be variable for the same question types. However, most theories of variability emphasise not just that participants will be variable in their response to different conjunction or disjunction problems but that they will also be variable for the same conjunction or disjunction problem if presented to them repeatedly. To investigate this, participants must provide repeated estimates for the same stimulus and their ‘internal’ variability must be examined in relation to their fallacy rates. This will allow us to examine whether variance for these responses will arise under these conditions.

3. Experiment 2

This experiment sought to examine the variance in probability estimates for simple natural language estimation tasks as in experiment 1. Here, we presented constituents, conjunctions and disjunctions repeatedly to participants asking them for estimates on different types of weather events. Noise accounts of cognitive biases emphasise that they result due to internal noise, that is, that a participant will have variable responses when they are asked for repeated estimates of the same event. In this experiment, each of the weather types were presented to participants repeatedly and in randomised order. Few experiments have looked at individual participant variability on the same probability judgements to date so repeated judgements will allow us to examine both variability and consistency of fallacy production for each participant.

3.1. Materials and method

The materials consisted of two sets of questions of the likelihood of specific weather conditions on a given day. The sets were designed so that participants were asked to assess weather conditions of high, medium and low likelihoods. Set A had four constituents (Windy, Sunny, Snowy, Cloudy), three conjunctions (Windy and Sunny, Windy and Snowy and Cloudy), and three disjunctions (Windy or Sunny, Windy or Cloudy, Snowy or Cloudy). Set B also consisted of four constituents (Warm, Rainy, Cold, Sunny), three conjunctions (Warm and Sunny, Rainy and Cold, Rainy and Warm), and three disjunctions (Warm or Sunny, Rainy or Cold, Rainy or Warm).

The questions about the likelihood of the weather conditions appeared on screen in front of the participants and they had to submit their estimate by moving a mark on a slider. Participants were asked What is the probability that the weather will be [weather type] on a randomly selected day in Ireland? The slider had a minimum value of 0 and a maximum value of 100. An estimate of 0 meant zero chance of that particular weather occurring on a given day. An estimate of 100 meant that the weather was certain to occur on a given day. To examine variability in estimates, each of the 10 set items were presented 5 times in a randomized order to the participant. In total, each participant was asked for 50 probability estimates. Unlike experiment 1, participants were only asked for probability responses. For this experiment, 87 participants were recruited from the student body in exchange for course credit. They were randomly assigned one of the two question sets. They were given a brief description of their task (assessing the likelihood of weather conditions on a given day) and informed that there was no time limit on task completion. The participants were asked to provide probability judgements for statements for the type of weather that appeared on-screen. At no stage did they have access to their previous responses.

3.2. Results

Participants who did not complete the task were excluded from the final analysis. In total, 6 participants failed to complete the task and were excluded from the final analysis. The results for the remaining 81 participants are given below.
3.2.1. Estimation and probability theory

We predict that participant estimates should be consistent with probability theory in terms of the addition law. Initially, averaged values for each of the A,B pairings were used to calculate the addition law itself. Overall, as with experiment 1, the participants’ estimates showed good compliance with the addition law. For all the pairings, the values were close to the expected normative value of 0, showing only mild deviations above and below that value, with an overall mean value of 0.004. Positive \((P(A) + P(B))\) and negative \((P(A \land B) - P(A \lor B))\) terms of the addition law were calculated for each participant (by averaging each participant’s 5 estimates for these terms for each pair A, B). The correlation between the pairs was \(r = 0.88, p < 0.00001\). Normative probability is represented by strongly correlated for the participants’ judgements, in all cases) and between average estimate difference and disjunction fallacy and a strong positive relationship between average estimate difference and conjunction fallacy rate, \(r = 0.86, p < 0.00001\). The difference for the disjunction \((A \lor B)\) terms of the addition law were calculated for each participant (by averaging each participant’s 5 estimates for these terms for each pair A, B). For the addition law to hold, the points must be symmetrically distributed around the line of identity. Fig. 2 shows the relationship between these positive and negative terms. A Deming regression was calculated using the participant estimates to investigate whether the estimates were consistent with probability theory. As in experiment 1, values for the addition law are distributed approximately symmetrically around the line of identity, with the line of best fit agreeing closely with the line of identity, as predicted by our model. A JZS Bayes Factor analysis based on a paired t-test of x and y values in this scatterplot gave strong evidence in favour of the null hypothesis that x and y values were equal (Scaled JZS Bayes Factor = 11.2), supporting the conclusion that the addition law identity holds in individual participant probability estimates.

3.2.2. Addition law and fallacy rate

The PTN predicts that the fallacies rates will be related via the addition law, with the conjunction fallacy rate relative to A following the disjunction fallacy rate relative to B, for any pairing of A,B. We see strong indications that this is the case in Table 5, with the fallacy rates for \(P(A \land B)\) vs \(P(A)\) and \(P(B)\) vs \(P(A \lor B)\) strongly correlated for the participants’ judgements, \(r = 0.841, p < 0.00001\).

3.2.3. Variability in probability estimation

The PTN predicts that the fallacy rates arise as a function of variance in the probability estimates with higher variance observable in the conjunction and disjunction statements than in the constituents. As each of the constituents, conjunctions and disjunctions were presented multiple times to participants, we were able to measure the variance for each event and type in the sample and examine how it relates to the observed fallacy rates. We tested this prediction, first by calculating the overall estimate difference for each conjunction and disjunction and comparing it to the fallacy rate for that item and then by comparing a restricted estimate difference to the fallacy rate. The overall estimate difference was calculated for each conjunction using \(P_D(A) - P_D(A \lor B)\). The difference for the conjunctions was calculated using \(P_D(A) - P_D(A \land B)\). This was then compared to the overall fallacy rate. A Pearson’s \(r\) correlation found a strong positive relationship between average estimate difference and conjunction fallacy rate, \(r = 0.862, p < 0.0005\) and a strong positive correlation between average estimate difference and disjunction fallacy rate, \(r = 0.86, p < 0.0005\).

As in experiment 1, to test whether these correlations held across different sets of participants, we performed 100 random split-half correlations, dividing participants into two randomly chosen equal-sized halves, calculating conjunctive and disjunctive fallacy rates for each pair of events A, B in one half and calculating estimate differences for those pairs in the other half, and measuring the correlation between those measures. There was strong positive relationship between average estimate difference and conjunction fallacy rate (average \(r = 0.83\), min \(r = 0.74\), \(p < 0.00001\) in all cases) and between average estimate difference and disjunction fallacy rate, (average \(r = 0.86\), min \(r = 0.75\), \(p < 0.00001\) in all cases).

Each participant in the experiment gave 5 probability estimates for each constituent, each conjunction, each disjunction, and so on. Individual conjunction and disjunction fallacy occurrences for a given constituent/conjunction pair were identified by comparing

Fig. 2. The figure above shows a scatterplot of all the positive and negative terms for all estimates and participants in experiment 2. Positive \((P(A) + P(B))\) and negative \((P(A \land B) - P(A \lor B))\) terms of the addition law were calculated for each participant (by averaging each participant’s 5 estimates for these terms for each pair A, B). The correlation between the pairs was \(r = 0.88, p < 0.00001\). Normative probability is represented by the line of identity, shown in grey. A Deming regression was calculated for both to determine the best fit line. This is represented by the dashed black line on the scatterplot.
these repeated estimates in order (if a participant’s first estimate for \(A \land B\) was greater than their first estimate for \(A\), a fallacy was recorded, if there second estimate for \(A \land B\) was greater than their second estimate for \(A\) a fallacy was recorded, and so on). To address the possibility that this ordering may have influenced responses we repeated the above split-half correlation test, but randomly shuffling the order of participants’ repeated estimates for each constituent event. Even with this random shuffling of repeated responses there remained a strong positive relationship between average estimate difference and conjunction fallacy rate (average \(r = 0.85\), min \(r = 0.70\), \(p < 0.00001\) in all cases) and between average estimate difference and disjunction fallacy rate, (average \(r = 0.85\), min \(r = 0.74\), \(p < 0.00001\) in all cases).

Finally, to address the fact that the fallacy rate and estimate difference are connected, the fallacy rate and estimate difference were separated and used to predict fallacy rates. The participants that had produced fallacies for a given conjunction or disjunction had these set of estimates excluded from the calculation of estimate differences. Then the difference \(P_A(A \land B) - P_A(A)\) was calculated for participants that didn’t produce a conjunction fallacy for a given conjunction. \(P_B(B) - P_B(A \lor B)\) was calculated for all the participants that didn’t produce a disjunction fallacy for a given disjunction. The fallacy rate was calculated for all instances of estimates in the pair. A significant positive correlation of \(r = 0.98\), \(p < 0.00001\) was observed for the restricted estimate difference and the conjunction fallacy rate. A significant positive correlation of \(r = 0.78\), \(p < 0.005\) was observed for the restricted estimate difference and the disjunction fallacy rate. For differences greater than 0, we see fallacy rates greater than 50%. For differences around 0, we see fallacy rates close to 50% and for differences less than 0, we see fallacy rates less than 50%. Table 5 displays the restricted differences and fallacy rates.

One of the predictions of the PTN model is that fallacy rate should occur inconsistently, that is - the participants should be variable in their responses-, when the calculated difference between the conjunction and constituent is zero. Inconsistent fallacy production occurred when the participant produced a fallacy response for 1, 2, 3 or 4 of the possible 5 occasions for each weather type. A consistent fallacy response occurred when the participants produced either 0 or 5 fallacy responses for the five possible occasions. For the sample, 51% of the responses were consistent and 49% were inconsistent. Each fallacy response and the corresponding average difference between the conjunction and constituent estimate were calculated. Fig. 3 shows the results of these calculations. Participants that produced zero fallacy responses (darkest frequency distribution, at the back of the graph) had an average difference between the conjunction and constituent estimates of zero or less. Participants that have five fallacy responses (lightest frequency distribution, at the back of the graph) had positive averages differences. Participants that had inconsistent responses have differences grouped around zero, with a pattern of increasingly positive results observed for the more fallacy responses made, just as predicted.

Variability. The total conjunction fallacy rate for the sample was 25%. A wide range of fallacy rates was observed for the conjunctions, with fallacy rates of 0% to 65% depending on the constituent-conjunction pair. An overall disjunction fallacy rate of 23% was observed for the sample. As with the conjunction fallacy rate, a wide range of fallacy rates were observed, here we observed fallacy rates of 0% to 63% depending on the constituent-disjunction pair. These results can be observed in Table 5. As in experiment 1, 95% confidence intervals were constructed for the constituents and complex items using the estimates where no fallacy had occurred. Then the overlap between the two CIs was compared to the fallacy rates. High fallacy rates typically occurred when there was both a positive overlap between the respective confidence intervals between the constituent and complex variability (see Tables 6, 7). A strong positive correlation was observed between the degree of CI overlap and the fallacy rate for both the conjunction

---

**Table 5**

| Constituent | Conjunction | Average Difference | Fallacy rate | Constituent | Disjunction | Average Difference | Fallacy rate |
|-------------|-------------|--------------------|--------------|-------------|-------------|--------------------|--------------|
| Cloudy Cloudy A Snowy | −0.54 | 0% | Snowy | Cloudy v Snowy | −0.50 | 0% |
| Windy Windy A Sunny | −0.53 | 0% | Sunny | Windy v Sunny | −0.30 | 0% |
| Rainy Rainy A Warm | −0.36 | 2% | Warm | Rainy v Warm | −0.27 | 5% |
| Cloudy Cloudy A Windy | −0.13 | 13% | Warm | Sunny v Warm | −0.13 | 10% |
| Sunny Sunny A Warm | −0.11 | 20% | Sunny | Sunny v Warm | −0.10 | 17% |
| Windy Cloudy A Windy | −0.11 | 25% | Windy | Cloudy v Windy | −0.18 | 20% |
| Warm Rainy A Warm | −0.12 | 27% | Cloudy | Cloudy v Windy | −0.12 | 20% |
| Cold Cold A Rainy | −0.10 | 27% | Rainy | Cold v Rainy | −0.16 | 24% |
| Warm Sunny A Warm | −0.08 | 27% | Cold | Cold v Sunny | −0.11 | 32% |
| Sunny Windy A Sunny | −0.07 | 45% | Windy | Windy v Sunny | −0.10 | 33% |
| Rainy Cold A Rainy | −0.11 | 49% | Rainy | Rainy v Sunny | −0.16 | 54% |
| Snowy Cloudy A Snowy | −0.01 | 65% | Cloudy | Cloudy v Snowy | −0.02 | 63% |
And the disjunction fallacy, \( r = 0.79, p < 0.005 \). The further apart the constituent and conjunction or disjunction values were, the less likely we were to observe a fallacy occurring, while fallacies were much more likely to occur where the confidence interval approached or exceeded zero.

**Individual variability.** As the participants had given multiple estimates for each constituent, conjunction and disjunction, we were able to assess each participant’s individual variability. In total, each participant has 6 occasions where we could compare the constituent and conjunction variability (e.g. Cloudy vs Cloudy\( \wedge \)Snowy and Snowy vs Cloudy\( \wedge \)Snowy) and 6 occasions where we could compare the constituent and disjunction variability (e.g. Cloudy vs Cloudy\( \wedge \)Snowy and Snowy vs Cloudy\( \wedge \)Snowy). For the conjunctions, 35% of participants were more variable for their individual constituent estimates than their conjunction estimates, the remaining 65% of participants were equally or more variable for their individual conjunction estimates. For the disjunctions, 30% of participants were more variable for their individual constituent estimates than their conjunction estimates, the remaining 70% of participants were equally or more variable for their individual disjunction estimates. The summary of these results can be seen in Fig. 4 where the individual variance is compared to the fallacy rates. Fallacies are more likely to occur when the conjunction or disjunction is more variable than the constituent.

### 3.3. Experiment 2 discussion

As with experiment 1, we investigated whether participant judgements were in agreement with probability theory. Again, we found strong evidence that their estimates were in line with the addition law, with only minor deviations being observed. While the estimates were consistent with this aspect of probability theory, participants still produced both conjunction and disjunction fallacies at varying rates, depending on the question posed to them.
To examine the variance in people's probability estimates for the same items, we repeatedly presented the participants with the same judgements and asked them to provide estimates for them on each occasion. This overwhelmingly demonstrated that participant judgments are noisy - estimates typically varied from one occasion to the next and the complex statements had more variance than the constituents. Constituents were typically more variable where the fallacy rate was close to 0%, the conjunctions were typically more variable where the high fallacy rates were observed. For disjunctions, they were usually more variable than their constituents regardless of fallacy rate. Additionally, high fallacy rates are commonly observed where the constituent and complex estimates are close to each other. This is consistent with the PTN which predicts that fallacy rates are due to the higher variance in the conjunction pushing the conjunction estimate above the constituent estimate. If the conjunction and constituent are close in value, then this is more likely to happen. Generally, it has been found in the literature that high probability - low probability constituent pairings (PLow(A) vs PLow(A)\textsuperscript{\lor}PHigh(B)) tend to produce the highest fallacy rates. Our results are consistent with this finding. However, here, as in the literature, “high-low” is a subjective observation of a constituent probability value that is a priori decided by the researcher rather than based on an objective, observable probability value. Further research is needed on whether this would hold if objectively high and low constituents were used.

Furthermore, we observed that participants are frequently inconsistent in producing fallacies for the same stimulus. For the conjunctions, nearly half of all the estimates were inconsistent i.e. participants produced a fallacy for some but not all of the repeated estimates for a given stimulus. A 100% fallacy rate for any of the stimuli was rare, with the majority of consistent responses being the 0% fallacy rate. If participants were producing their estimates using a heuristic-based approach, we would expect to see participants consistently producing or avoiding a fallacy for the repeated estimates.

Analysis of the relationship between fallacy rates and variance in probability estimates demonstrated that fallacies typically occurred where the conjunction or disjunction was more variable than the constituent. Complex statements are typically more variable than constituents. Fallacy rates are a product of both variance in the estimates and the “true” probability values of the constituents.

### Table 6
Confidence Intervals for conjunction estimates (Exp 2). As in experiment 1, the restricted estimates were used to calculate the 95% confidence intervals. A positive value for the overlap meant that the estimates were typically close to each other, while a negative overlap meant that the estimates were typically far from each other. A reliable positive correlation was observed between fallacy rate and confidence interval overlap, \( r = 0.71, p < 0.01 \).

| P(A)       | P(A \& B)         | P(A \& B) 95% CI | P(A \& B) 95% CI | Overlap | Fallacy |
|------------|-------------------|------------------|------------------|---------|---------|
| Cloudy     | Cloudy\textsuperscript{\&}Snowy | 0.64 (0.73) | 0.08 (0.15) | -0.49 | 0% |
| Windy      | Windy\textsuperscript{\&}Sunny  | 0.58 (0.69) | 0.35 (0.45) | -0.13 | 0% |
| Rainy      | Rainy\textsuperscript{\&}Warm   | 0.53 (0.65) | 0.18 (0.27) | -0.26 | 2% |
| Cloudy     | Windy\textsuperscript{\&}Cloudy | 0.65 (0.75) | 0.51 (0.62) | -0.03 | 13% |
| Sunny      | Warm\textsuperscript{\&}Sunny   | 0.31 (0.40) | 0.20 (0.28) | -0.03 | 20% |
| Windy      | Windy\textsuperscript{\&}Cloudy | 0.60 (0.72) | 0.49 (0.61) | 0.01  | 25% |
| Warm       | Warm\textsuperscript{\&}Sunny   | 0.27 (0.36) | 0.19 (0.28) | 0.01  | 27% |
| Cold       | Rainy\textsuperscript{\&}Cold   | 0.56 (0.70) | 0.45 (0.60) | 0.04  | 27% |
| Warm       | Rainy\textsuperscript{\&}Warm   | 0.25 (0.35) | 0.14 (0.23) | -0.02 | 27% |
| Sunny      | Windy\textsuperscript{\&}Sunny  | 0.35 (0.46) | 0.28 (0.39) | 0.04  | 45% |
| Rainy      | Rainy\textsuperscript{\&}Cold   | 0.51 (0.66) | 0.40 (0.56) | 0.05  | 49% |
| Snowy      | Cloudy\textsuperscript{\&}Snowy | 0.02 (0.06) | 0.02 (0.05) | 0.03  | 65% |

### Table 7
Confidence Intervals for Disjunction Estimates (Exp 2). The 95% confidence intervals for the constituent-disjunction pairs for experiment 2. A positive overlap demonstrated that the estimates for the constituent-disjunction pair overlapped to some degree. A strong positive correlation is observed between the CI overlap and fallacy rate, \( r = 0.791, p < 0.005 \). The greater the negative overlap, the lower the fallacy rate. Positive overlaps are associated with fallacy rates greater than 50%.

| P(B)       | P(A \textsuperscript{\lor} B) | P(A \textsuperscript{\lor} B) 95% CI | Overlap | Fallacy |
|------------|-------------------------------|--------------------------------------|---------|---------|
| Snowy      | Cloudy\textsuperscript{\&}Snowy | 0.05 (0.09) | 0.51 (0.62) | -0.42 | 0% |
| Sunny      | Windy\textsuperscript{\&}Sunny | 0.34 (0.42) | 0.64 (0.72) | -0.22 | 0% |
| Warm       | Rainy\textsuperscript{\&}Warm | 0.26 (0.34) | 0.52 (0.63) | -0.18 | 5% |
| Warm       | Warm\textsuperscript{\&}Sunny  | 0.26 (0.33) | 0.38 (0.48) | -0.05 | 10% |
| Sunny      | Warm\textsuperscript{\&}Sunny  | 0.28 (0.37) | 0.38 (0.48) | -0.01 | 17% |
| Windy      | Windy\textsuperscript{\&}Cloudy | 0.55 (0.68) | 0.75 (0.84) | -0.07 | 20% |
| Cloudy     | Windy\textsuperscript{\&}Cloudy | 0.62 (0.72) | 0.75 (0.83) | -0.03 | 20% |
| Rainy      | Rainy\textsuperscript{\&}Cold  | 0.50 (0.63) | 0.67 (0.77) | -0.04 | 24% |
| Cold       | Rainy\textsuperscript{\&}Cold  | 0.52 (0.64) | 0.65 (0.75) | -0.01 | 32% |
| Windy      | Windy\textsuperscript{\&}Sunny | 0.52 (0.65) | 0.63 (0.74) | 0.02  | 33% |
| Rainy      | Rainy\textsuperscript{\&}Warm  | 0.39 (0.56) | 0.57 (0.71) | -0.01 | 54% |
| Cloudy     | Cloudy\textsuperscript{\&}Snowy | 0.53 (0.70) | 0.55 (0.72) | 0.15  | 63% |
stimuli. Little research has been done where the objective probability is known or even available to the researchers. Stimuli such as the “Linda problem” have only subjective probabilities. For other research that uses real world events - like the weather events used here or future sporting events (e.g. Teigen, Martinussen, & Lund, 1996) - it might be possible to calculate objective values for them but at best these are non-stationary and it is typically truer to say that these have subjective probability values too. To fully understand the role of probability in producing estimates and its impact on fallacy rates and whether participants are objectively skilled reasoners, participants must produce estimates for stimuli that have accessible, objective probability values. In situations where participants produce fallacies at a varying rates but where we have no access to objective probabilities, we cannot fully determine why this range of results exists. We investigate this in experiment 3.

4. Experiment 3

For experiment 3, we investigated the impact of two factors on variability: probability values and sample size. Typically, research on the conjunction and disjunction fallacy has not used experimental stimuli that has observable objective probabilities: most of the research to date has employed stimuli with subjective probabilities - e.g. scenarios about people. The binomial model predicts that variance in estimates is a result of the probability values of the stimuli. Here, we present the participants with simple judgements where the underlying, objective probability is controlled. This will allow us to examine how variability of estimates relates to probability value. In addition to this, we will also look at the role of sample size in probability estimates and variability. To this end, participants will be presented with stimuli that preserves the underlying probabilities while modifying the sample size.

To examine the internal variability of the participants, we present them with repeated probability judgements. They saw images where each image contains a set number of shapes differing in colour (red, white or green) and configuration (solid or hollow). For each image, participants were asked to estimate the probability of an event (a randomly selected shape being red, for example). The true probability of events in these images were held constant across multiple presentations (but with the images themselves varying as to the position of the shapes on the screen each time), as described below. Each participant saw multiple presentations of the same probability question (multiple questions for which the objectively correct probability was the same), allowing us to estimate the degree of random variation in participant estimates. Some questions asked about simple events (a shape being red, being hollow, etc.) while other questions asked about conjunctive and disjunctive events (a shape being red and solid, a shape being white or hollow, etc.) Two distinct sets of images were used, with objective probabilities held constant in each set (see below). The images from these two sets were interspersed with each other. Participants answered questions about 460 images in total. Images were only on screen for a short time (2 s), so participants did not have time to count the occurrence of shapes of different types. Images were presented in randomised order.

4.1. Materials

The images consisted of shapes of three colours - colours C1, C2, and C3 respectively - and 2 shape configurations - S1 and S2 - with fixed probabilities. To prevent the participants from remembering or recognising the images after multiple repetitions, the actual
colour varied from image to image, so sometimes colour C1 was white, sometimes colour C1 was red and sometimes colour C1 was green but the objective probability value assigned to C1 remained the same. The colours varied in the same way for colour C2 and colour C3. The actual configuration of the shapes also varied from image to image so sometimes configuration S1 was the solid shapes and sometimes configuration S1 was the hollow shapes. As with C1, the objective probabilities were held constant. Conjunction and disjunctions were created for a number of combinations of colour and configuration such as \(P(C_1 \land S_1)\), \(P(C_1 \land S_2)\) and \(P(C_2 \lor S_1)\).

For each type \(C_1, C_2, C_3, S_1, S_2, C_1 \land S_1, C_1 \land S_2\) etc, there were 20 images asking participants to estimate the probability of that type. In practise, this meant that the participants saw 20 images asking them to estimate the probability of colour C1, 20 images asking them to estimate the probability of colour C2, 20 images asking them to estimate the probability of configuration S1 and so on.

Each image presentation included a question to elicit a probability judgement. For the colour probability questions, participants were presented with questions in the form “What is the probability of picking a shape that is [colour C1]?” or “What is the probability of picking a shape that is [colour C2]?” For the configuration questions, participants were presented with questions in the form: “What is the probability of picking a shape that is [configuration S1]?” or “What is the probability of picking a shape that is [configuration S2]?”

The conjunction and disjunction questions took the same form. For instance, the question to elicit a probability judgement for the disjunction of C1 and C2 would be: “What is the probability of picking a shape that is [colour C1 AND configuration S1]?”.

### 4.1.1. Set 1 - Probability values

The stimuli in set 1 was designed to investigate how probability values effect variability. In set 1, colour C1 had a fixed probability of 0.7, colour C2 had a fixed probability of 0.2 and colour C3 had a fixed probability of 0.1. Configuration S1 had a fixed probability of 0.9 and configuration S2 had a fixed probability of 0.1. The conjunctions for set 1 were created using the following colour and configuration combinations: \(P(C_1 \land S_1)\), \(P(C_2 \land S_1)\), and \(P(C_2 \land S_2)\). These corresponded to the objective probability values of 0.63, 0.18 and 0.02. The disjunctions for set 1 were created using the following colour and configuration combinations: \(P(C_1 \lor S_1)\), \(P(C_2 \lor S_1)\), \(P(C_2 \lor S_2)\), and \(P(C_3 \lor S_2)\). These disjunction combinations corresponded to the objective probability values of 0.97, 0.92, 0.73, and 0.28. Participants viewed 220 images of 20 geometric shapes on a computer screen with a probability question. Colour C3 was excluded from the probability questions for this set (as both C1 and S2 had objective probability value of 0.1).

### 4.1.2. Set 2 - Sample size

Set 2 was designed to investigate how sample size effects probability estimation and variability. To this end, the probability values were fixed for each level. For set 2, colour C1 had the fixed probability of 0.333, colour C2 had the fixed probability of 0.333 and colour C3 had the fixed probability of 0.333. Configuration S1 had the fixed probability value of 0.5 and configuration S2 had the fixed probability value of 0.5. The conjunction for set 2 had the value of 0.17. Any combination of \(C_1, C_2, C_3\) and \(S_1, S_2\) would give this value. The disjunction had the objective probability value of 0.67. Again, any combination of \(C_1, C_2, C_3\) or \(S_1, S_2\) would give this value.

Participants viewed 240 images of geometric shapes in a computer screen. Each image consisted of 12, 24, or 36 shapes (levels 1, 2, and 3 respectively). Each objective probability values of 0.333, 0.5, 0.17, and 0.67 were presented 20 times for each of the 12, 24 and 36 shape images. At the bottom of each image was a question asking participants about the probability of some event (shape, color, or shape/color conjunction) given the sample shown in the image. This was followed by a slider scale: participants moved the bar on this scale to select their estimated probability for the event in question. A box to the right showed the currently selected probability paired with a button labelled ‘next’: clicking that button recorded the participant’s probability estimate and moved the participant on to the next screen (see Fig. 5). For ease of use the slider’s position remained where the participant had placed it as the participant moved on to the next screen.

### 4.2. Procedure

Participants were seated at a screen. Each participant began with a training trial of sample stimuli to familiarize themselves with the task. Training trials used different probability combinations to the main experiment. Once the participants were comfortable with the task, they moved onto the experimental trials.

The static image and the probability question appeared on screen simultaneously. The image was replaced with a blank screen once 2 s had elapsed to prevent the participants from counting the shapes. The associated question remained on-screen until the participants had made their guess. The participants indicated their estimate by moving a mark on a slider using their mouse or arrow keys. This slider had a minimum value of 0 and a maximum value of 1. Responses were discretized. A box in the corner indicated the exact value of the participants’ estimate and dynamically updated as they moved the slider. When the participant was satisfied with their answer, they submitted it by clicking on a “Next” button. This also triggered the succeeding image and probability question.

### 4.3. Results

A total of 9 participants made 460 probability judgements each. Their responses and response time was recorded for each judgement. Two of the participants were excluded from the final analysis for failing to answer over 20% of the questions. The number of participants is consistent with other studies of probability perception (e.g. Gallistel, Krishan, Liu, Miller, & Latham, 2014).

#### 4.3.1. Estimation and probability theory

To test whether there is evidence of normative reasoning in the participants’ estimates, we employed the addition law in a variety of ways. The estimates for \(A, B\), and their conjunction and disjunction combinations were used to calculate the addition law values as
in the previous experiments. For estimates to be compliant with probability theory, the terms should cancel to zero. The addition law
was calculated for estimates in both set 1 and set 2. In set 1, the addition law was calculated for the following:

\[
P(C_1) + P(S_1) - P(C_1 \land S_1) - P(C_1 \lor S_1) = 0
\]

\[
P(C_2) + P(S_2) - P(C_2 \land S_2) - P(C_2 \lor S_2) = 0
\]

\[
P(C_2) + P(S_1) - P(C_2 \land S_1) - P(C_2 \lor S_1) = 0
\]

Consistent with the previous experiments, the identities were close to zero, varying minutely around the value. An overall value of 0.037 was found for the sample. Fig. 6 shows the scatterplot with the positive and negative terms for these addition law identities in participant responses. A Deming regression was calculated using the participant estimates to investigate whether the estimates agreed with the addition law prediction. As in experiment 1, values for the addition law are distributed approximately symmetrically around the line of identity, with the line of best agreeing closely with the line of identity, as predicted by our model. A JZS Bayes Factor analysis based on a paired t-test of x and y values in this scatterplot gave strong evidence in favour of the null hypothesis that x and y values were equal (Scaled JZS Bayes Factor = 25.5), again confirming the conclusion that the addition law identity holds in individual participant probability estimates.

4.3.2. Addition law and fallacy rate

The PTN predicts that the fallacy rates will be related to the addition law. The conjunction fallacy rate relative to A should follow the disjunction fallacy rate relative to B, for any pairing of A,B. This arises as a natural consequence of the addition law. We see strong indications that this is the case, with the fallacy rates for \( P(A \land B) \) vs \( P(A) \) and \( P(B) \) vs \( P(A \lor B) \) strongly correlated for the participants’ judgements in set 1, \( r = 0.91, p < 0.0005 \) and set 2, \( r = 0.832, p < 0.001 \). This can be observed in Table 8.

4.3.3. Estimate accuracy

Representativeness accounts typically posit that participants are poor estimators of probability. Here, we can investigate how accurate judgements are by comparing them to the objective probability. For each of the 11 probability values in set 1, each participant gave 20 estimates for its value. In set 2, the 4 probability estimates were questioned at 3 different levels; 20 estimates were given for each probability value at each level. The relationship between mean probability estimates and objective probability are displayed in Fig. 7.

For each probability value, the participants’ average estimate and standard deviation were calculated. The average estimate and standard deviation were also calculated for the sample. The average deviation from the true probability was calculated in terms of percentage points. Some noticeable trends were observed, participants tended to overestimate the low probabilities and underestimate the higher probabilities. The degree of overestimation for the low constituents was much less than for the low complex statements. For instance, the constituent with a true probability of 0.1 have an average estimate of 0.13, while the conjunction with a probability of 0.02 had an average estimate of 0.14.

Overall, conjunctions were overestimated and disjunctions were underestimated. The conjunctions’ average deviation from their objective value was by 10 percentage points, the disjunction average deviation their true probability was by 17 percentage points while the constituent average deviation was by 7 percentage points. Fig. 7 shows the average estimate for each type. For set 2, the conjunctions were overestimated on all occasions, with the average estimate increasing as the stimulus set became more complex.

\[^4\text{No estimates were elicited for } P(C_1 \land S_1) \text{ so the addition rule could not be calculated for combinations of } P(C_1) \text{ and } P(S_1).\]
The disjunctions were consistently underestimated. Participants were more accurate in their estimates for the constituents. The 12-shape combinations had the lowest average estimates, the 24-shape estimates were higher than the 12-shape and lower than the 36-shape estimates. The 36-shape images had the highest mean estimates.

4.3.4. Variability in probability estimation

As with the previous experiments, the average estimate difference for the complex item and its constituents were calculated and compared with the fallacy rate for that item. Significant positive correlations were observed for both the conjunction average difference and fallacy rate, $r = 0.66$, $p < 0.05$, and the disjunction average difference and fallacy rate, $r = 0.73$, $p < 0.01$. A consistent relationship was observed between the average difference and the fallacy rate where higher fallacy rates are associated with positive average

### Table 8

| Set 1 | $P(A \land B)$ | $P(A)$ | Average Difference | Fallacy rate | $P(A)$ | $P(A \land B)$ | Average Difference | Fallacy rate |
|-------|-----------------|--------|--------------------|-------------|--------|-----------------|--------------------|-------------|
| 0.9   | 0.18            | -0.70  | 0%                 | 0.2         | 0.92   | 0.62            | 0%                 | 0.2         |
| 0.2   | 0.02            | -0.10  | 14%                | 0.1         | 0.28   | -0.14           | 6%                 | 0.1         |
| 0.9   | 0.63            | -0.11  | 15%                | 0.7         | 0.97   | -0.12           | 17%                | 0.7         |
| -     | -               | -      | -                  | 0.7         | 0.73   | -0.11           | 18%                | 0.7         |
| 0.2   | 0.18            | -0.10  | 19%                | 0.2         | 0.28   | -0.11           | 36%                | 0.2         |
| 0.1   | 0.02            | -0.04  | 42%                | 0.9         | 0.92   | -0.09           | 45%                | 0.9         |
| 0.7   | 0.63            | -0.06  | 68%                | 0.9         | 0.97   | -0.08           | 71%                | 0.9         |

| Set 2 | $P(A \land B)$ | $P(A)$ | Average Difference | Fallacy rate | $P(A)$ | $P(A \land B)$ | Average Difference | Fallacy rate |
|-------|-----------------|--------|--------------------|-------------|--------|-----------------|--------------------|-------------|
| 0.5†  | 0.17            | -0.27  | 4%                 | 0.33‡       | 0.67   | -0.23           | 14%                | 0.67‡       |
| 0.5§§ | 0.17            | -0.23  | 4%                 | 0.33‡       | 0.67   | -0.2            | 20%                | 0.67§§      |
| 0.5‡  | 0.17            | -0.22  | 11%                | 0.33‡       | 0.67   | -0.23           | 22%                | 0.67‡       |
| 0.33§ | 0.17            | -0.19  | 13%                | 0.5§§       | 0.67   | -0.15           | 23%                | 0.67§§      |
| 0.33§ | 0.17            | -0.16  | 19%                | 0.5§§       | 0.67   | -0.15           | 24%                | 0.67§§      |

Note: †12 shapes, ‡24 shapes, § 36 shapes

The disjunctions were consistently underestimated. Participants were more accurate in their estimates for the constituents. The 12-shape combinations had the lowest average estimates, the 24-shape estimates were higher than the 12-shape and lower than the 36-shape estimates. The 36-shape images had the highest mean estimates.

![Fig. 6](image-url) The figure above shows a scatterplot of all the positive and negative terms for all estimates and participants in experiment 3. The positive term is the sum of the $P(A)$ and $P(B)$ estimates while the negative term is the sum of the $P(A \land B)$ and $P(A \lor B)$ estimates. The correlation between the pairs for set 1 was $r = 0.794$, $p < 0.00001$. The pairs for set 2 had a correlation value of $r = 0.79$, $p < 0.00001$. As the groups were very similar in their estimates, they were collapsed. For the scatterplot, normative probability is represented by the line of identity, shown in grey. A Deming regression was calculated to determine the best fit line. This is shown in black on the scatterplot. For the addition law to hold, the points must be symmetrically distributed around this line of identity.
differences and lower fallacy rates are associated with negative differences. As before, the restricted estimate difference was calculated and used to predict fallacy rates. The fallacy rate was calculated for each pair and any instance where a participant had made the fallacy was excluded from the analysis of the average difference. The results can be observed in Table 8. There was a significant positive correlation between the restricted estimate difference and fallacy rate for conjunctions, $r = 0.57$, $p = 0.05$ and for disjunctions, $r = 0.63$, $p < 0.05$.

Each conjunction and constituent was presented 20 times to each participant. To evaluate the rate at which the participant had committed the conjunction fallacy, each conjunction judgement 1…20 was matched in order with its corresponding constituent judgements 1…20, so the first conjunction judgement was matched with the first constituent judgements, and so on. If a particular conjunction judgement exceeded the estimate of either of the corresponding constituent values, an instance of the conjunction fallacy was recorded. For each participant, there were six conjunction questions where the fallacy could be committed, three from set 1 and three from set 2. The average conjunction fallacy rate was 19%. Fallacy rates ranged from 0% to 68% per constituent-conjunction pair: a range that is in-line with those seen in description based studies (e.g. Stolarz-Fantino et al., 2003). The set-up of this experiment allows us to categorise conjunctions based on their actual probabilities and their underlying constituent probabilities. The participants showed marked differences in performances for each of the six conjunctions they were presented with. Table 8 displays the fallacy rate breakdown by conjunction type.

As with the conjunction fallacy, each disjunction judgement was matched with the constituent judgements in sequence, so the first disjunction judgement was matched with the first instances of the relevant constituent judgements. If a disjunctive estimate was less than either of its constituent estimates then it was counted as an instance of the disjunction fallacy. The average disjunction fallacy rate was 24%. The fallacy rate ranged from 0% to 71%, which is consistent with the results from description based research and simulations of the PTN. The average fallacy rate for the each of the 7 possible disjunctions is displayed Table 8. As for the conjunctions, the objective probability value of the disjunction was not an indicator of fallacy rate occurrence. Conjunction and disjunction fallacy occurrence varied over the course of presentation, however, there was no obvious trend of improvement or deterioration in the participants ability to avoid committing the fallacies (that is, fallacy rates did not decline with task familiarity).

In experiment 2, participants gave 5 repeated estimates of the same probability question so we could measure the internal variance and the consistency of fallacy production of each participant. In experiment 3, participants gave 20 estimates for each objective probability. Here, an inconsistent response occurred where the participant produced a fallacy on 1–19 of the possible occasions for a conjunction or disjunction. A consistent fallacy response occurred when the participant produced a fallacy response on 0 or 20 of the occasions. The fallacy response rates were calculated for each participant in addition to the average difference in estimate between the conjunction and constituent. These results are displayed in Fig. 8. Participants with low fallacy rates typically had negative differences in estimates, with increasingly positive estimate differences as the fallacy rate rose. The maximum number of fallacies committed by any of the participants was 17 (of a possible 20). In total, 27% of the fallacy responses were consistent, with a participant either producing a fallacy in all responses for a given item, or in no responses for that item (all the consistent responses involved no fallacy production) and 73% of the responses were inconsistent (with the same participant sometimes producing fallacy responses for a given item and sometimes not).

**Variance.** Since each conjunction, disjunction and constituent was presented 20 times to each participant, we can estimate the degree of variance (standard deviation) in estimates for type. Recall that the PTN model predicts greater variance would exist for the complex combinations than the constituents. The average SDs revealed that the conjunctions were noisier than their constituent counterparts for 75% of the comparisons. In a breakdown by participant, the conjunctions were more variable on 33% of the occasions to 75% of the occasions, depending on the participant. The average SDs for the disjunctions showed that they were more
variable than their constituent counterparts for 100% of the comparisons. This ranged from 64% of the occasions to 100% of the occasions, depending on the participant.

This supports the PTN model assumption that conjunction and disjunction fallacies arise due to variability in conjunction and disjunction estimates. Overall, the complex combinations had higher average standard deviations than the constituents.

Overall, Levene’s test found statistical significance in 62% of the comparisons. The conjunctions were found to have statistically higher levels of variance on 17% of the occasions and the constituents being statistically more variable than the conjunctions on 8% of occasions. For the disjunctions, Levene’s test found that they had significantly higher levels of variance than their constituents in 93% of the comparisons.

95% confidence intervals were constructed for each conjunction - constituent and disjunction - constituent pair. The overlap between the confidence intervals, the difference in their respective SDs and the fallacy rate can be seen in Tables 9 and 10. A significant positive correlation was observed for the CI overlap and the conjunction fallacy rate, $r = 0.598$, $p < 0.05$. A significant positive correlation was also observed for the CI overlap and the disjunction fallacy rate, $r = 0.65$, $p < 0.05$. Higher fallacy rates were observed where the constituent and complex statement were close to each other.

**Individual Variability.** As in experiment 2, the nature of experiment 3 - repeated elicitation of probability estimates - allows us to examine individual variability in participant estimates. In total, there were 12 constituent and conjunction comparisons for each participant - e.g. $P_C(C_1) \times P_C(C_1 \land S_1)$. There were 14 occasions where the constituent and disjunction variability could be compared, e.g. $P_C(C_1) \times P_C(C_1 \lor S_2)$. The relationship between variance and fallacy rates for individual participants is displayed in Fig. 9. Typically, higher fallacy rates were associated with positive differences in SD - that is, that the complex statement was more variable than the constituent.

**4.4. Experiment 3 discussion**

As we observed in experiments 1 and 2, the participant estimates for this experiment were consistent with the addition law of probability theory, showing only the mild deviations observed in the other experiments. Despite the different stimuli used compared
to the previous experiments (estimates for language statements vs estimates for visual stimuli), the participants still produced estimates consistent with this aspect of normative reasoning. From this, we can assume that the reasoning process employed in both scenarios are consistent and the results are comparable. The set-up for this experiment is relatively novel in work on cognitive biases and to our knowledge, only a small number of studies (e.g. Wedell & Moro, 2008) have investigated conjunction or disjunction fallacies where the underlying probability was known and none have explicitly looked at variance in those responses. It gives us a position to examine how participant estimates relate to objective estimates, how fallacy rates are influenced by probability values and how sample size effects estimation. Overall, participants were most accurate for the constituents and more accurate for the conjunctions than disjunctions.

Fallacy rates observed here (0–68% for conjunctions, 0–71% for disjunctions) are in line with those observed for other conjunction/disjunction studies. We addressed the observation in the literature that the “high-low” constituent pairing of 0.9, 0.2, only produced fallacy rates of 0% and 19% respectively. In fact, the difference between the constituent was to the conjunction or disjunction, the higher the resulting fallacy rate was likely to be.

As in the experiments with the description based stimuli, average estimate difference was a good predictor of fallacy rate, with positive correlations observed for both the conjunction and disjunctions. However, this brings into question the exact role of probability values in fallacy responses. If average difference and variance are a very good predictors of fallacy rates then it is possible

Table 9
Confidence Intervals for Conjunction estimates (Exp 3). The 95% confidence intervals for the constituent-conjunction pairs in experiment 3 based on the restricted estimates. A positive value for the overlap meant that the estimates were typically close to each other, while a negative overlap meant that the estimates were typically far from each other. A positive correlation between the restricted CI overlap and fallacy rate was observed, \( r = 0.598, p < 0.05 \).

| P(A) | P(A ∩ B) | Low | High | Overlap | Fallacy |
|------|---------|-----|------|---------|---------|
| 0.9  | 0.18    | 0.84| 0.87 | −0.67   | 0%      |
| 0.51 | 0.17    | 0.46| 0.49 | −0.24   | 4%      |
| 0.55 | 0.17    | 0.48| 0.5   | −0.20   | 4%      |
| 0.55 | 0.17    | 0.43| 0.46 | −0.25   | 11%     |
| 0.33 | 0.17    | 0.39| 0.42 | −0.23   | 13%     |
| 0.33 | 0.17    | 0.41| 0.45 | −0.26   | 13%     |
| 0.5  | 0.63    | 0.87| 0.89 | −0.78   | 15%     |
| 0.33 | 0.17    | 0.32| 0.36 | −0.19   | 19%     |
| 0.2  | 0.18    | 0.22| 0.25 | −0.14   | 19%     |
| 0.1  | 0.02    | 0.13| 0.16 | −0.02   | 42%     |
| 0.7  | 0.63    | 0.75| 0.8  | −0.01   | 68%     |

Note: †12 shapes, ‡24 shapes, § 36 shapes.

Table 10
Confidence Intervals for Disjunction Estimates (Exp 3). The 95% confidence intervals for the constituent-disjunction pairs for experiment 3. The restricted participant estimates were used to calculate the confidence intervals. The larger the negative overlap, the smaller the fallacy was likely to be, as the overlap got closer to 0, the fallacy rate increased. A reliable positive correlation is observed between the restricted CI overlap and fallacy rate, \( r = 0.65, p < 0.05 \).

| Constituent | Disjunction | Low | High | Overlap | Fallacy |
|-------------|-------------|-----|------|---------|---------|
| 0.2         | 0.92        | 0.21| 0.23 | −0.58   | 0%      |
| 0.1         | 0.73        | 0.12| 0.14 | −0.64   | 0%      |
| 0.1         | 0.28        | 0.12| 0.14 | −0.11   | 6%      |
| 0.33†       | 0.67        | 0.3 | 0.33 | −0.19   | 14%     |
| 0.7         | 0.97        | 0.69| 0.73 | −0.09   | 17%     |
| 0.7         | 0.73        | 0.69| 0.73 | −0.08   | 18%     |
| 0.33§       | 0.67        | 0.37| 0.4  | −0.16   | 20%     |
| 0.33‡       | 0.67        | 0.37| 0.41 | −0.18   | 22%     |
| 0.5§        | 0.67        | 0.4 | 0.44 | −0.11   | 23%     |
| 0.5‡        | 0.67        | 0.46| 0.49 | −0.12   | 24%     |
| 0.2         | 0.28        | 0.18| 0.2  | −0.07   | 36%     |
| 0.5†        | 0.67        | 0.42| 0.47 | −0.11   | 40%     |
| 0.9         | 0.92        | 0.81| 0.86 | −0.05   | 45%     |
| 0.9         | 0.97        | 0.75| 0.83 | −0.03   | 71%     |

Note: †12 shapes, §24 shapes, § 36 shapes.
that probability values play no direct role in fallacy rates - rather it could be due entirely to higher variance in the complex item and the absolute differences between the two values that fallacy rates occur. The PTN predicts that higher fallacies will occur with conjunctions close to 0.5 than at the extremes. Currently, however, we cannot say conclusively that this is the case. In the following experiment, we address this by controlling the distance between constituents and conjunctions.

As the objective probabilities were available for this experiment, we were able to directly test the values predicted by the binomial model against those we calculated from the participant estimates. Overall, we observed that the model was consistent with the participant data, with the participants’ variance showing the same trends that the binomial model predicts. Under the binomial model, the variance in an estimate is predicated by its probability value and a conjunction fallacy response occurring is most likely where \( P(A \land B) = 0.5 \) and the \( P(A) \) value is further from 0.5 than the conjunction and so not as variable as the conjunction. The results suggest a pattern that is consistent with the predictions of the binomial variance model. However, the \( p \) values here are not evenly spread across the scale, nor are we able to directly compared the variance (as predicted by the binomial model) for \( P(A) \), \( P(A \land B) \) and \( P(A \lor B) \) responses of the same \( p \) value, so we cannot draw strong conclusions about the accuracy of the model. We address these issues in the following experiment.

5. Experiment 4

In experiment 3, we examined the role of probability value in estimate variance. Here, we take that further to look at the variability of constituents and conjunctions with the same objective probability value. We aim to examine how making different probability judgements effects the response while the objective probability remains the same. To this end, we will compare the probability estimates and variance for constituents and conjunctions of the same objective probability value. It is expected that greater variance and greater deviation from the objective value will be observed for the conjunctions.

The previous experiments have looked at estimate difference as a predictor of fallacy rate. The average difference between the constituent and conjunction has been shown to be a good indicator of fallacy rate. Here, we control the distance between the constituent and conjunction, so that they are either \(|0.1|\) or \(|0.15|\) apart. The PTN argues that probability values will impact fallacy rates. If participants are accurate in their judgement, then we would expect to see differences of approximately \(-0.1\) or \(-0.15\) (depending on the pair) between the constituent-conjunction pairs \( (P(A \land B) - P(A)) \), and we would expect to see low fallacy rates and minimal difference between the fallacy rates for each of the pairings.

Again, this experiment involves repeatedly presenting participants with images where each image contains a set number of shapes differing in colour and configuration. For each image participants are asked to estimate the probability of some event (e.g., a randomly selected shape being red). The true probability of events in these images were held constant across multiple presentations. Each participant saw multiple presentations of images for which the objectively correct probability was the same, allowing us to estimate the degree of random variation in participant estimates. Some questions asked about simple events (a shape being red, being hollow, etc.) while other questions asked about conjunctive events (a shape being red and solid, etc.) Images were only on screen for a short time (2s), so participants did not have time to count the occurrence of shapes of different types. Images were presented in randomised order.

Fig. 9. This graph shows the relationship between the difference in variance for individual estimates and the fallacy rate for materials in experiment 3. Each participant gave multiple estimates for the same constituent, conjunction and disjunction, so individual fallacy rates and variance for probability estimates could be calculated for each participant. Fallacies typically occurred when there was a positive overlap in confidence intervals and when there was a positive difference in variance - that is, when the complex item was more variable than the constituent. Low fallacy rates were more likely to occur when there was a negative difference in variance or no overlap between constituent and complex CI.
5.1. Materials

The material set for this experiment consisted of 192 images, each with 20 shapes of varying types and colours. The images were organised into 7 ‘probability sets’ so that all images in a given set contained the same number of occurrences of some constituent \( A \), the same number of occurrences of some constituent \( B \), and the same number of occurrences of the conjunction \( A \land B \). These event counts (and hence the probability objectives of these events \( A, B, \text{ and } A \land B \)) were the same in all images in a given set: these counts and probabilities are given in Table 11. However, the actual concrete instantiation of each event varied randomly from image to image within each set (so that in one image in the first set, \( A \) would be represented by red, \( B \) by solid, and there would be 5 red shapes 5 solid shapes, and 3 hollow red shapes; while in another image in the same set \( A \) would be represented by hollow and \( B \) by blue and there would be 5 hollow shapes, 5 blue shapes, and 3 hollow blue shapes; and so on). The position of shapes also varied randomly across images. This variation in event representation and position was designed to ensure that participants could not respond by recalling estimates given for previous images: all images were unique.

There were 24 images for each probability set - 12 presented with a question asking participants to estimate the probability of single event \( A \) (however it was represented in that particular image) and 12 presented with a question asking participants to estimate the probability of conjunctive \( A \land B \) (however it was represented in that particular image). In addition to these 7 probability sets there was a filler set containing 12 images with single-event questions and 12 with conjunctive event questions, but with no relation between those single and conjunctive events. The full set of images were presented in random order: images were not grouped according to probability set, and filler images were interspersed throughout.

Probability sets were designed so that probabilities presented to participants would be either 0.15, 0.25, 0.35, 0.5, 0.65, 0.75, 0.85 or 0.95, for both single event \( A \) and conjunctive event \( A \land B \). This was to allow direct comparison of single and conjunctive probability estimates for cases where the single event and the conjunctive event had the same underlying objective probability. Participants were only asked to estimate probabilities for event \( A \) and event \( A \land B \) in each set: estimates for event \( B \) were not obtained.

As in Experiment 3, each image was paired with a question asking participants about the probability of some event (shape, color, or shape/color conjunction) given the sample shown in the image. This was followed by a slider scale: participants moved the bar on this scale to select their estimated probability for the event in question. A box to the right showed the currently selected probability paired with a button labelled ‘next’: clicking that button recorded the participant’s probability estimate and moved the participant on to the next screen (see Fig. 5). In this experiment the slider’s position was reset to the center of the probability scale when the participant moved on to the next screen.

5.1.1. Procedure

Participants were seated at a screen. They began with a training trial of sample stimuli to familiarize themselves with the task. Once the participants were comfortable with the task, they moved onto the experimental trials. The static image appeared on screen. The image was replaced with a black screen with a fixation point once 2 s had elapsed to prevent the participants from counting the shapes. The probability question appeared once the static experimental image has disappeared. The question remained on-screen until the participants had made their estimate. The participants indicated their estimate by moving a slider using their mouse. This slider had a minimum value of 0 and a maximum value of 1. A box in the corner indicated the exact value of the participants’ estimate and dynamically updated as they moved the slider. When the participant was satisfied with their answer, they submitted it by clicking on a “Next” button. This also triggered the succeeding image and probability question.

5.2. Results

In total, 12 participants produced estimates for 192 images. Both their responses and response time were recorded. The results are detailed below. For each conjunction judgement \( P(A \land B) \), it had its own constituent, \( P(A) \), against which we could check

| \( A \) | \( B \) | \( A \land B \) | \( P_B(A) \) | \( P_E(A) \) | \( P_B(B) \) | \( P_E(B) \) | \( P_B(A \land B) \) | \( P_E(A \land B) \) |
|---|---|---|---|---|---|---|---|---|
| 5 | 5 | 3 | 0.25 | 0.323 | 0.25 | – | 0.15 | 0.253 |
| 7 | 7 | 5 | 0.35 | 0.387 | 0.35 | – | 0.25 | 0.330 |
| 10 | 10 | 7 | 0.50 | 0.479 | 0.50 | – | 0.35 | 0.421 |
| 13 | 13 | 10 | 0.65 | 0.580 | 0.65 | – | 0.50 | 0.481 |
| 15 | 15 | 13 | 0.75 | 0.705 | 0.75 | – | 0.65 | 0.604 |
| 17 | 17 | 15 | 0.85 | 0.766 | 0.85 | – | 0.75 | 0.714 |
| 19 | 19 | 17 | 0.95 | 0.873 | 0.85 | – | 0.85 | 0.785 |
conjunction fallacy rates. In addition to this, it had a value matched constituent \( P(C) \), which had the same objective probability as the conjunction \( P(A \land B) \) but the conjunction was not a subset of the constituent \( C \), so no fallacy rates could be derived from their comparison.

5.2.1. Estimate accuracy

As in experiment 3, we were able to compare the subjective responses to objective population probability values. For each of the 8 objective values, there was a constituent and a conjunction response elicited for its value. The relationship between the average probability estimates and the objective “true” probability values is displayed in Fig. 10. As each objective value has both constituent and conjunction responses, we are able to examine the role that type plays in probability estimation. Typically, the following trends were observed, regardless of type: probability estimates were overestimated for values less than 0.5, estimates for the objective value of 0.5 were the most accurate and estimates for values about 0.5 were underestimated. Fig. 10 also displays the average amount of deviation from the true probability value. Constituents had less deviation from the true probability value than the conjunctions for values less than 0.5, similar amounts of deviation for 0.5, and greater deviation from the objective value for values above 0.5. Conjunctions had greater deviation from the objective probability for values less than 0.5, the same deviation for 0.5, and less deviation from the objective probability value for values over 0.5.

5.2.2. Variability in probability estimation

As expected, the total conjunction fallacy rate for the sample was very low, with an average of 24%. As the objective difference was controlled between the constituents and conjunctions, it was hypothesised that there would be no relationship between average difference and fallacy rates. Pearson’s correlation found no significant relationship, \( r = 0.405, p > 0.05 \). The fallacy rate and average estimate difference were partitioned and calculated as in the previous experiments. A relationship was not observed for the restricted estimate differences and fallacy rate and no correlation was found between the two, \( r = -0.655, p > 0.05 \).

As with experiment 2 and 3, each participant saw multiple presentations of each item which allowed us to test the PTN prediction that participants will produce the fallacy in an inconsistent fashion for the same item. For this experiment, fallacy rates of 0 or 12 (of a possible 12) were counted as a consistent fallacy response while the occurrence of 1–11 (of a possible 12) fallacies per item were counted as an inconsistent fallacy response. The majority of responses were inconsistent and no participant had 100% fallacy rate for any of the conjunctions. The maximum observed fallacy rate by any participant for any of the conjunctions was 75% (9 out of a possible 12). Fig. 11 displays the fallacy rate occurrence and its corresponding average estimate difference. In total, 13% of the fallacy responses were consistent (no participant produced 12 fallacy responses for any of the conjunctions so all the consistent responses are 0 fallacies here) and 87% of the responses were inconsistent. Participants that produced zero fallacy responses had average values grouped around zero with increasingly positive results as the rate of fallacy production increased.

**Group Variance.** With this experiment, we could examine variance in two ways; variance between the constituent-conjunction pairings \( P(A) \) vs \( P(A \land B) \) and variance between the value-matched constituents and conjunctions \( P(C) \) vs \( P(A \land B) \).

Overall, the conjunctions \( P(A \land B) \) were more variable than their constituents \( P(A) \) on 86% of occasions. Levene’s test of equality
of variances found that the conjunctions were statistically significantly more variant than the constituents on 72% of the occasions. In addition to this, we also matched the constituents $P(C)$ and conjunctions $P(A \wedge B)$ that had the same objective probability values to compare how type effects variance. For example, the constituent with the objective value of 0.35 was matched to the conjunction that had the objective value of 0.35 and their response variances were found. Overall, the conjunctions were more variable on 75% of the occasions compared to the value-matched constituents. Again, Levene’s test was used to determine if any of the conjunctions were statistically significantly more variable than the value matched constituents. In this case, statistical significance was found for 38% of the comparisons.

Binomial Variance. Using the participants’ measured variance for each probability estimate, we tested the predicted variance values from the binomial model. To predict the variance, we assume that $K$ (the number of successes) is distributed according to the binomial distribution, $K \sim \text{Bin}(N, p)$. As we are interested in the sample proportion, $K/N$, rather than the sample count, $K$, we calculate the variance using

$$\frac{p(1-p)}{N}$$

where $N = 12$, the number of repetitions of each item. This model predicts that the highest variance will be observed for estimates $X$ where $P(X) = 0.5$ and the variance should decline the closer the estimates are to $P(X) = 0$ or $P(X) = 1$. Each participants’ variance was calculated and compared to the predicted value. Fig. 12 displays the measured and predicted variance versus the objective probability. The participant values are distributed around the predicted value in all cases, with lower variance typically found close to 0 and 1 and high variance found close to the midpoint. The variance values closely follows the predictions of the binomial model. We observe that participants typically had low variance where the model predicted low variance and high variance where the model predicted high variance and that the model predictions are a good fit for the data. Polynomial fits were calculated for both the constituents and conjunctions and found good fits for both against the predicted values. The measured individual variances for items was positively correlated with the predicted variance for that item, $r = 0.51, p < 0.00001$. Observed variance in people’s probability estimates for both constituents and conjunctions followed the variance values predicted by the binomial model with no observable
difference between them. A t-test found no significant difference between the constituent and conjunction variance, $t(95) = -1.156$, $p > 0.05$.

### 5.2.3. Aggregate-level model fitting

This experiment, involving as it does objective probability values for $P(A)$ and $P(A \land B)$, allows us to test the computational fit between our model’s predicted means and standard deviations (SDs) for probability estimates $P_E(\Lambda)$ and $P_E(A \land B)$ and the means and SDs of participant responses. To carry out this fit we select values for the noise parameters $\delta$ and $\delta \Delta$ (used to calculate predicted mean probability estimates $P_E(\Lambda)$ and $P_E(A \land B)$) for given objective probabilities $P(A)$ and $P(A \land B)$, as in Eqs. (1) and (2)) and for the sample size parameter $N$ (used to calculate predicted variance for these probability estimates, as in Eq. (6): predicted SD is the square root of this variance). Prior to fitting we can identify a reasonable range of values for these free parameters. We expect the noise rate $d$ to be relatively low (somewhere around 0.1), the best fitting value in previous computational fits of this model: see Costello & Watts (2017)), and we expect parameter $\delta \Delta$ to be significantly smaller than this value $d$. Finally, we expect the sample size parameter $N$ to be somewhere around to Miller’s ‘magical number’ $7 \pm 2$’ for working memory capacity (Miller, 1956).

We take the best fit between model and data to occur when the Root Mean Squared Difference (RMSD) between predicted and observed mean probability values, and predicted and observed SDs, is minimised. These values were minimised for parameter values $d = 0.1, \delta \Delta = 0.02, N = 7$. With these parameters the RMSD between participants’ mean probability estimates and predicted mean estimates (computed from objective probability values as in Eqs. (1) and (2)) was $RMSD = 0.021$ (correlation between observed and predicted values, $r = 0.994, p < 0.00002$, across all single and conjunctive events; for single events alone these parameters gave a fit of $RMSD = 0.021, r = 0.994$; for conjunctive events alone these parameters gave a fit of $RMSD = 0.022, r = 0.995$). With these parameters the RMSD between average SD in participants’ probability estimates and predicted SD for estimates for those events (computed from objective probability values by taking the square root of the value in Eq. (6)) was $RMSD = 0.017$ (correlation between observed and predicted SD, $r = 0.73, p < 0.05$, across all single and conjunctive events; for single events alone these parameters gave a fit of $RMSD = 0.017, r = 0.76$; for conjunctive events alone these parameters gave a fit of $RMSD = 0.009, r = 0.899$). The model is a good fit to people’s average probability estimates, and standard deviation in those estimates, for a reasonable set of parameter values.

We can also fit the Nilsson et al. (2009) configural weighting model of conjunctive probability estimation, as given in Eq. (7), to this experimental data and compare with the fit given by the PTN model. Since the configural weighting model does not address constituent probability estimation (it simply assumes such estimates are available, but does not explain how they are produced) and does not make any specific predictions about variance in estimates, this fit can examine only conjunctive probability estimates. Note that the probability sets used in this experiment were designed so that both constituents $P(A)$ and $P(B)$ had the same objective probability for the first six probability sets: in these sets $P_E(A)$ and $P_E(B)$ are expected to be equal, and so the configural weighting model predicts that $P_E(A \land B) = P_E(A) = P_E(B)$ should hold irrespective of weighting parameter $W$ in these cases. Since this weighting parameter $W$ affects only the value of $P_E(A \land B)$ in the 7th probability set, a value for $W$ was chosen so that the averaging model exactly matched participants mean conjunctive probability estimate for that set (see Table 12). As this table shows, estimates produced by the PTN were closer to those produced by participants (with lower RMSD, higher correlation $r$) than those produced by the averaging model, though the correlation between averaging model conjunctive estimates and participants’ average conjunctive estimates was also very high ($r = 0.986$, versus $r = 0.995$ for the PTN model). This high correlation produced by the averaging model
could well be an artefact of the experimental design. The averaging model predicts that \( P_E(A \land B) \) will equal \( P_E(A) \) (for the first 6 probability sets), but these materials were specifically designed so that the objective probability \( P_E(A \land B) \) exactly followed the objective probability \( P_E(A) \) (this design being used to test differences in variance for single and conjunctive events with the same objective probability). This designed-in relationship could explain the observed high correlation between \( P_E(A \land B) \) and the averaging model’s predicted value of \( P_E(A) \). We investigate and compare model fits further in the next section, where we describe computational fits at the individual level to participant’s repeated estimates in Experiments 2, 3, and 4, and carry out model comparisons based on those fits using WAIC.

5.3. Experiment 4 discussion

In experiment 4, we investigated how judgement type effects probability estimates by presenting participants with constituents, \( P(C) \), and conjunctions, \( P(A \land B) \), of the same value and elicited repeated responses for each of them. Typically, we see that participants are good at estimating both types of judgements - with only marginal differences in mean estimates for a given objective probability value. The most accurate estimates for both constituents and conjunctions were for \( P = 0.5 \), while estimates for items where \( p \) was less than 0.5 were overestimated and those above 0.5 were underestimated. This pattern is consistent with the PTN where noise has a regressive effect towards 0.5, causing estimates below 0.5 to be overestimated and estimates above 0.5 to be underestimated. In both experiment 3 and 4, we have observed that participants are typically accurate reasoners, particularly for constituents and conjunctions but we have also observed that they frequently produce inconsistent responses to the same conjunction stimulus, sometimes committing the fallacy and sometimes avoiding it entirely. The participant estimates were accurate for both constituents and conjunctions, however, the conjunctions were more variable than the constituents in both the case of the \( P(A) \) and \( P(C) \) judgements. Aggregate-level model fitting demonstrates that our model’s predictions are good fits for both the participants’ mean and SDs. These fits were also performed for the configural weighted model. While our model performed better, the configural model also produced strong correlations between the model fits and data. In the following section, we investigate the models performances for fits at an individual level.

6. Individual level computational model fitting

In this section we describe computational model fits to individual participant responses in Experiments 2, 3 and 4 for both the binomial variance and the configural weighting models. The model fitting process was carried out in Stan, a probabilistic programming language that provides full Bayesian statistical inference with MCMC sampling (Carpenter et al., 2017). Model fitting was carried out simultaneously on individual (repeated) probability estimates for constituents, conjunctions, and disjunctions and on (conjunctive and disjunctive) fallacy occurrence for those individual estimates. The same general framework was used for both models. We first consider model fitting for experiments 3 and 4 (for which objective probabilities of events are known), and then consider fitting for experiment 2 (for which objective probabilities of events are not known and must be treated as free parameters in the model-fitting process).

6.1. Binomial variance model fitting

In fitting the binomial variance model to individual repeated probability estimates in a given experiment with known objective probabilities of events (Experiments 3 and 4), we assume 3 free parameters for each participant: \( d_{\text{implic.}} < 0.5 \) (the noise rate for that
participant in estimating probabilities for simple events A and B, assumed to be less than 0.5), $d_{\text{complex},i} < 0.5$ (the noise rate for that participant in estimating complex events $P(A \land B)$ and $P(A \lor B)$, assumed to be less than 0.5) and $N_i$ (the sample size used by that participant in estimating probabilities). Given the known objective probability for some event $A$ we assume that participant $i$’s repeated probability estimates for that event are randomly distributed around the mean estimate

$$P_i(A) = (1 - 2d_{\text{simple},i})P(A) + d_{\text{simple},i}$$

(where $P(A)$ is the objective probability for event $A$), with standard deviation

$$\sigma_{A,i} = \sqrt{\frac{P(A)(1 - P(A))}{N_i}}$$

For modelling purposes we assume the error distribution around the mean estimate is approximately normal: so that participant $i$’s repeated estimates for event $A$ follow the Normal distribution

$$\text{Normal}(P_i(A), \sigma_{A,i})$$

(8)

For complex events $A \land B$ (or $A \lor B$) we assume that participant $i$’s repeated probability estimates are randomly distributed around the mean estimates

$$P_i(A \land B) = (1 - 2d_{\text{complex},i})P(A \land B) + d_{\text{complex},i}$$

(9)

$$P_i(A \lor B) = (1 - 2d_{\text{complex},i})P(A \lor B) + d_{\text{complex},i}$$

(10)

(where $P(A \land B)$ and $P(A \lor B)$ are the known objective probabilities for those events) with standard deviations

$$\sigma_{A \land B,i} = \sqrt{\frac{P_i(A \land B)(1 - P_i(A \land B))}{N_i}}$$

$$\sigma_{A \lor B,i} = \sqrt{\frac{P_i(A \lor B)(1 - P_i(A \lor B))}{N_i}}$$

and participant $i$’s repeated estimates for events $A \land B$ and $A \lor B$ follow the Normal distributions

$$\text{Normal}(P_i(A \land B), \sigma_{A \land B,i})$$

(11)

$$\text{Normal}(P_i(A \lor B), \sigma_{A \lor B,i})$$

(12)

The binomial variance model assumes that the relationship $0 \leq d_{\text{simple}} \leq d_{\text{complex}} \leq 0.5$ holds for these noise parameters (all noise parameters are less than 0.5, and noise for simple events is less than noise for complex events). We implement this in our model by defining two free parameters $0 \leq d_{\text{simple},i} < 0.5$ and $0 \leq d_{\text{increase},i} < 1$ for each participant, and, given this, calculate $d_{\text{complex},i}$ as

$$d_{\text{complex},i} = d_{\text{simple},i} + (0.5 - d_{\text{simple},i})d_{\text{increase},i}$$

so that every possible value of the free parameters $d_{\text{simple},i}$ and $d_{\text{increase},i}$ will produce a value of $d_{\text{complex},i}$ such that $0 \leq d_{\text{simple}} \leq d_{\text{complex}} \leq 0.5$ as required.

The binomial variance model fit depends on the objective probability for simple, conjunctive and disjunctive events. In Experiment 2, these objective probabilities are not known. In fitting to experiment 2, therefore, we augment the model with additional free parameters representing the objective probabilities of these events (which we assume are common across all experimental participants). These free parameters representing (unknown) objective probabilities are constructed to be fully consistent with all normative requirements of probability theory. Recall that Experiment 2 contained two separate sets of events, each containing 4 single events, 3 conjunctions, and 3 disjunctions (9 objective probabilities in total). For each set 7 free parameters were required to construct these 9 objective probabilities and ensure full consistency with probability theory.

To test the binomial variance model’s prediction that $d_{\text{simple}} \leq d_{\text{complex}}$ hold, we also carry out secondary fits with a version of the model that simply treats the noise rates $d_{\text{simple},i}$ and $d_{\text{complex},i}$ as independent parameters that can take on any value (less than 0.5). This secondary fit allows us to test the binomial variance model predictions about differential noise rates by comparing degree of fit for constrained ($d_{\text{simple}} \leq d_{\text{complex}}$) and unconstrained ($d_{\text{simple}}$ and $d_{\text{complex}}$ independent) versions of the model.

6.2. Configural weighting model fitting

In fitting the configural weighting model to individual repeated probability estimates, we use the same approach of modelling error in repeated probability estimates as normally distributed around the mean estimate produced by the model. We assume 3 free parameters for each participant $i$: $\sigma_{\text{simple},i}$ (the standard deviation of that participants repeated probability estimates for simple events $A$ and $B$), $\sigma_{\text{complex},i}$ (the standard deviation of that participants repeated probability estimates complex events $A \land B$ and $A \lor B$) and $W_i$ (the participant’s weighting parameter used in calculating complex estimates from the configural weighting of simple probabilities). In experiments 3 and 4, where objective probabilities for events are known, we assume that participant $i$’s repeated probability estimates for some simple event $A$ follow the Normal distribution

$$\text{Normal}(P_i(A), \sigma_{\text{simple},i})$$

(13)
where \( P_i(A) \) is the mean of participant \( i \)'s probability estimates for that event. Note that the configural weighting model doesn’t give any account of the relationship between constituent probability estimates and objective probability values. To fit the configural weighting model to Experiments 3 and 4, we assume that the mean constituent estimate for a given participant, \( P_i(A) \), is a linear function of the true probability of event \( A \):

\[
P_i(A) = O_i(1 - S_i) + S_i \times P(A)
\]

where the ‘scale’ parameter \( 0 \leq S_i \leq 1 \) represents participant \( i \)'s mapping from the objective probability scale to their own subjective estimate scale, and the offset parameter \( 0 \leq O_i \leq 1 \) represents the intercept of that mapping (multiplied by \( (1 - S_i) \) to ensure all constituent probability estimates fall between 0 and 1).

For complex events \( A \land B \) we assume that participant \( i \)'s repeated probability estimates are randomly distributed around the mean estimates

\[
P_i(A \land B) = W_i \min(P_i(A), P_i(B)) + (1 - W_i) \max(P_i(A), P_i(B)) \quad 0.5 \leq W_i \leq 1
\]

and

\[
P_i(A \lor B) = (1 - W_i) \min(P_i(A), P_i(B)) + W_i \max(P_i(A), P_i(B)) \quad 0.5 \leq W_i \leq 1
\]

given in the configural weighting model, and that they follow the Normal distributions

\[
\text{Normal}(P_i(A \land B), \sigma_{\text{complex},i})
\]

(14)

\[
\text{Normal}(P_i(A \lor B), \sigma_{\text{complex},i})
\]

(15)

For Experiment 2, where objective probabilities of events are not known, we fitted the configural weighted model in a way that matched the fitting approach for the binomial variance model: by adding free parameters to represent mean probability estimates for single events. Experiment 2 contained two separate sets each containing 4 single events, and so we fitted the configural weighting model by adding 4 additional free parameters for each set. These single probabilities could take on any value between 0 and 1.

6.3. Fitting individual conjunction and disjunction fallacy responses

As well as fitting participant’s repeated probability estimates for single and conjunctive/disjunctive events, we are also interested in fitting conjunction and disjunction fallacy responses in those estimates. Both the binomial variance and the configural weighting models see conjunction fallacy rates as a function of the difference of means \( P_i(A \land B) - P_i(A) \), and of random variation or noise in estimates.

In fitting the binomial variance model we assume that individual probability estimates follow the normal distributions given in Eqs. (8), (11) and (12). This means that the difference between estimates for a constituent \( A \) and a conjunction \( A \land B \), for participant \( i \), will follow the distribution

\[
\text{Normal}(P_i(A \land B) - P_i(A), \sigma_{\text{simple},i} + \sigma_{\text{complex},i})
\]

Given this, the probability of a conjunction fallacy for these events in participant \( i \)'s responses is equal to the probability of obtaining a positive value under this distribution; and this probability is given by

\[
P_i(A \land B > A) = 1 - \Phi(0; P_i(A \land B) - P_i(A), \sigma_{\text{simple},i} + \sigma_{\text{complex},i})
\]

(16)

Where \( \Phi \) is the cumulative function for this normal distribution. Similarly, the probability of a disjunction fallacy occurring is given by

\[
P_i(A \lor B < B) = 1 - \Phi(0; P_i(A \lor B) - P_i(B), \sigma_{\text{simple},i} + \sigma_{\text{complex},i})
\]

(17)

In fitting the configural weighting model we similarly assume that individual probability estimates follow normal distributions given in Eqs. 8, 11 and 12

\[
\text{Normal}(P_i(A \land B) - P_i(A), \sigma_{\text{simple},i} + \sigma_{\text{complex},i})
\]

Given this, the probability of a conjunction fallacy occurring is equal to the probability of obtaining a positive value under this distribution; and this probability is given by

\[
P_i(A \land B > A) = 1 - \Phi(0; P_i(A \land B) - P_i(A), \sigma_{\text{simple},i} + \sigma_{\text{complex},i})
\]

(18)

and the probability of a disjunction fallacy occurring is given by

\[
P_i(A \lor B < B) = 1 - \Phi(0; P_i(A \lor B) - P_i(B), \sigma_{\text{simple},i} + \sigma_{\text{complex},i})
\]

(19)

Since in a given item the conjunction fallacy either occurs or does not occur (it is a binary variable), and since the chance of occurrence is a function of the difference \( P_i(A \land B) - P_i(A) \), a natural distributional model for fallacy occurrence is the Bernoulli distribution. In our computational fit for both models, therefore, we represent the distribution of conjunction fallacy occurrences in repeated estimates for events \( A \land B \) and \( A \) produced by a given participant \( i \) as

\[
\text{Bernoulli}(P_i(A \land B > A))
\]

and the distribution of disjunction fallacy occurrences in repeated estimates for events \( A \lor B \) and \( B \) produced by a given participant \( i \) as
We thus have a common framework for computational fits of both the binomial variance and configural weighting models to experimental data. In this framework repeated individual probability estimates are modelled as normally distributed around a mean computed by the model (via the regressive or noisy means with parameters $d_{\text{simple},i}$ and $d_{\text{complex},i}$ in the binomial variance model; via weighting of constituent probabilities with parameter $W_i$ in the configural model) with a given standard deviation (calculated from the mean and sample size parameter $N_i$ in the binomial variance model; taken as free parameters $\alpha_{\text{simple},i}$ and $\alpha_{\text{complex},i}$ in the configural model), while occurrence/non-occurrence of the conjunction and disjunction fallacies are modelled via the Bernoulli distribution parameterised as described above. In this framework the binomial variance model has three free parameters $d_{\text{simple},i}$, $d_{\text{complex},i}$ and $N_i$ for each participant, while the configural model has five: $\alpha_{\text{simple},i}$, $\alpha_{\text{complex},i}$, $W_i$, scale parameter $S_i$ and intercept parameter $O_i$. We fit these models to experimental data using Stan, a probabilistic programming language for specifying statistical models that provides full Bayesian inference for continuous-variable models using a adaptive form of Hamiltonian Monte Carlo sampling (Carpenter et al., 2017). Stan probabilistic programs implementing these models for Experiments 2, 3 and 4, along with raw experimental data and R code running computational fits of these models, are available online.\textsuperscript{5} We compared model fit using the Widely Applicable Information Criterion (WAIC, Watanabe, 2010) which takes functional form complexity into account when comparing model fit, implemented in R (Vehtari, Gelman, & Gabry, 2017; Vehtari, Gelman, & Gabry, 2018). We first describe the results of fitting these models to results from Experiment’s 3 and 4, for which objective probabilities for events are known; we then describe model fits to results from Experiment 2, for which objective probabilities are not known.

7. Computational fit results

We used Stan to implement the binomial variance and configural weighting models as described above and applied them to participant’s individual responses in Experiment 2 (set A and set B), Experiment 3 and Experiment 4. Experiment 2 asked 40 participants (set A) or 41 participants (set B) to give 5 repeated probability estimates for 4 simple constituent events, 3 conjunctions, and 3 disjunctions, giving 2000 individual estimates in set A and 2050 estimates in set B. Experiment 3 asked 7 participants to give 20 repeated probability estimates for three sets of events, with each set containing two constituents, one conjunction, and one disjunction, and for one set of events containing 4 constituents, 3 conjunctions and 4 disjunctions, giving 3220 individual estimates in total. Due to a coding error and some network problems, a relatively small number of these responses were not recorded (99 dropped responses out of 3220 or 3% of responses dropped, distributed randomly across all response sets). Since Stan does not handle missing data, we cleaned the set of raw response data by replacing any blank responses in a given participant i’s repeated estimates for some event A with the average of the remaining responses given by participant i for that event A. Finally, Experiment 4 asked 12 participants to give 12 repeated probability estimates for 7 constituents and 7 conjunctions, giving 2016 individual estimates in total; in this experiment 9 responses were blank (due to network problems); these were replaced with that participants average estimate for the event in question, as before.

After preparing the data we fitted the two models to each experimental dataset via the stanfit MCMC sampler called from R, with 4 chains and 2000 iterations per chain. Fits for both models converged ($\hat{R} = 1$) in all cases except when the configural weighted model was applied to the two sets of data from Experiment 2 (even much larger numbers of iterations, up to 20,000, did not produce convergent fit for the configural weighted model on this datasets; given that the binomial variance model converged for these sets, this suggests that the configural weighted model is a poor model for the data in Experiment 2). In all cases the fit was to both individual probability estimates and to individual conjunction and disjunction fallacy occurrences (treated as categorical data as described above). For Experiment 2 we ran a single fit for each model. For experiments 3 and 4 we ran two fits. On the first run we extracted log likelihoods, and hence WAIC expected log pointwise predictive densities ($\text{elpd}_{\text{WAIC}}$), for the two models across all individual constituent, conjunction and disjunction probability estimates and all individual conjunctive and disjunctive fallacy occurrences. On the second run we extracted log likelihoods and $\text{elpd}_{\text{WAIC}}$ values for the two models for individual conjunction and disjunction probability estimates and all individual conjunctive and disjunctive fallacy occurrences (dropping constituent probability estimates because both models fit these estimates very closely). Table 13 gives $\text{elpd}_{\text{WAIC}}$ values, differences $\text{elpd}_{\text{DIFF}}$, and standard errors, for both runs. Note that difference in extracted log likelihood values does not affect the fit produced, only the data returned for a given fit.

Table 13 shows WAIC expected log predictive density adjusted for number of parameters ($\text{elpd}_{\text{WAIC}}$) for the two models in each dataset, alongside expected log predictive density difference ($\text{elpd}_{\text{DIFF}}$) between the two models, and standard errors for these values. Higher log predictive densities indicate better fits; negative values of $\text{elpd}_{\text{DIFF}}$ indicate preference for the first model (binomial variance): the binomial variance gave a better fit in all cases. Since $\text{elpd}_{\text{DIFF}}$ values are approximately normal we use the Z test to indicate statistically significant differences in model fit. The binomial variance model had a statistically significant advantage in model fit (at $p < 0.05$ or lower) for Experiment 2 and Experiment 4. For experiment 3 there was no significant difference in model fit.

7.1. Fitting to individual participants separately

The above analysis compares model fits across all individual participants and responses, and shows an overall advantage for the binomial variance model. The above fits implicitly assume that all participants follow the same process of probability estimation, and

\textsuperscript{5} https://osf.io/a47ut/.
Table 13

| Target values                        | binomial variance | configural weighting | elpdWAIC (SE) |
|--------------------------------------|-------------------|----------------------|---------------|
| Expt 2 set A                         | 115.0 (116.4)     | -615.6 (93.7)        | -730.6 (172.9)†|
| Expt 2 set B                         | -582.7 (99.0)     | -820.1 (88.1)        | -237.5 (70.0)†|
| Expt 3 all                           | 241.0 (261.8)     | 219.3 (266.1)        | 21.8 (109.4)  |
| Expt 3 excluding single events       | -1016.4 (198.1)   | -1089.5 (192.7)      | 72.1 (104.3)  |
| Expt 4 all                           | 249.6 (113.0)     | 169.6 (114.4)        | -80.0 (42.4)‡ |
| Expt 4 excluding single events       | -219.3 (87.0)     | -319.6 (90.4)        | 100.3 (32.2)† |

Note: †: p < 0.001, ‡: p < 0.05.

asks whether that process is better modelled by the binomial variance or the configural weighting approach. It could be argued that different participants might follow different approaches to probability estimation, with some participants following one approach and some the other. To test this proposal, we fit the two models to each individual participant’s responses for all target values (excluding constituent estimates, which were easy to fit for both models) in Experiments 3 and 4, and compared model fits for each participant. We did not carry out this process of fitting models separately to participants in Experiment 2, primarily because of difficulties with convergence for the configural weighting model in that experiment. Table 14 shows the results of these separate fits. Testing for statistical significance of difference in fit at the p < 0.05 level (with Bonferroni correction for multiple comparisons, giving operational criteria of significance of p < 0.0026) we found that fits were indistinguishable most participants, but significantly in favour of the binomial variance model in two cases.

7.2. Conjunction and disjunction fallacy rate predictions

To assess the level of agreement between observed conjunction and disjunction fallacy rates and rates predicted by the two models, we extracted observed conjunction and disjunction fallacy rates for each participant and each conjunction/constituent and disjunction/constituent pairing in all Experiments. We also extracted values for the expressions in Eqs. (16) and (17) (the binomial variance model’s predicted fallacy rates) and for the expressions in Eqs. (18) and (19) (the configural weighting model’s predicted fallacy rates) from the model fits described above. We then calculated the correlation between observed and predicted fallacy rates for the two models (see Table 15). For Experiment 4, for example, where there were 12 participants and 12 conjunction/constituent pairs, these numbers represent the correlation between 12 × 12 = 144 observed conjunction fallacy rates (each participant’s repeated estimates for each conjunction/constituent pair producing an observed fallacy rate for that participant and that pair, equal to the proportion of times that participant gave a higher estimate for the conjunction than the constituent in those repeated estimates) and 144 predicted fallacy rates produced by the model in question. All correlations were positive and significant at the p < 0.01 level: correlations produced by the binomial variance model were higher than those produced by the configural weighting model in all cases. Both models tended to overestimate fallacy rates for both conjunctions and disjunctions (see the positive differences between predicted and observed fallacy rates in the Table) but the binomial variance model’s predicted fallacy rates were closer to the observed rate in all cases.

7.3. Relation between d_simple and d_complex

The binomial variance model predicts the noise rate for complex events should be higher than the noise rate for simple events: that d_simple < d_complex. The model fits described above impose this requirement on the noise parameters explicitly. To illustrate the relationship between these noise rates in the model, Fig. 13 shows a scatterplot of best-fitting d_simple versus d_complex values for the 81 participants in Experiment 2. As the figure shows, the difference was noticeable across a range of participants.

To test the model’s prediction that d_simple < d_complex, we reran the model fits described above, but with the binomial variance model modified so that the requirement d_simple < d_complex was not imposed: in this version of the binomial variance model both d_simple and d_complex could independently take on any value between 0 and 0.5. We compared the model fit obtained with this unconstrained version of the model against the fit obtained with the constrained (d_simple < d_complex) version. If degree of fit to data for this unconstrained version of the model was noticeably greater than that obtained from the constrained version where d_simple < d_complex, that would count as evidence against the model’s prediction. Note that we do not expect the constrained version of the model to give a significantly better fit than the unconstrained version, since the unconstrained version can always ‘find’ parameter values for which d_simple < d_complex, happens to hold (and so can match the degree of fit of the constrained model). Table 16 shows the expected log predictive densities for these two versions of the model across experiments 2, 3 and 4: the two versions are essentially indistinguishable, which supports the binomial variance account.
The aim of the paper was to examine variability in probability estimation and its relationship to two well known cognitive biases - the conjunction and disjunction fallacies. To this end, we carried out four experiments; the first a study of variance and how different response formats effects probability judgements, the second was a study of the internal variance in probability estimation which was tested by giving the participants repeated judgement tasks. Both these experiments used description-based stimuli consistent with other description-based studies of the conjunction and disjunction fallacy. The third experiment focused on the roles of probability values and sample size on variance, estimation accuracy and fallacy rates. The final experiment again looked at the role of probability values on fallacy rates and how question type influences estimation. Experiments 3 and 4 both employed repeated judgements to understand the internal variance in participant estimates and the effect of $p$ values on that variance. Each had stimuli with observable objective probabilities so we could investigate estimate accuracy. This makes these experiments somewhat novel in research on cognitive biases. However, fallacy rates observed for both were in line with more traditional research stimuli in this field so we believe that they are appropriate.

Results showed that variability of the estimate is a key indicator of whether a fallacy response will occur. Overall, the complex statements showed higher levels of variability with statistically significant levels observed for most occasions. Approximately 70% of the complex statements were more variable than their constituent counterparts across all experiments. A small number of the simple

### Table 14
WAIC expected log predictive density adjusted for number of parameters (elpdWAIC) for the binomial variance and configural weighting model fit separately each individual participant’s responses in Experiments 3 and 4, for all target values excluding constituent estimates (which were easy to fit for both models). Negative elpdWAIC values indicate preference for the first model (binomial variance). Fits were indistinguishable for all participants but two, for which the fit comparison was significantly in favour of the binomial variance model (at Bonferroni corrected $p < 0.05$).

| Participant | binomial variance | configural weighting | elpdWAIC (SE) | elpdWAIC (SE) |
|-------------|-------------------|----------------------|---------------|---------------|
|             |                   |                      |               |               |
| 1           | −67.8 (87.3)      | −176.4 (68.1)        | −108.6 (52.8) |
| 2           | −59.7 (88.7)      | −110.9 (83.4)        | −51.2 (51.2)  |
| 3           | −193.3 (63.8)     | −149.9 (72.5)        | 43.4 (35.8)   |
| 4           | −213.1 (61.9)     | −202.4 (65.3)        | 10.7 (16.5)   |
| 5           | −131.1 (74.0)     | −167.6 (72.3)        | −36.5 (20.9)  |
| 6           | −198.9 (70.4)     | −145.0 (75.2)        | 54.0 (40.2)   |
| 7           | −152.9 (75.5)     | −138.0 (76.3)        | 14.9 (39.0)   |

### Table 15
Each column shows the correlation, $r$, between observed fallacy rates (from experimental data) and predicted fallacy rates extracted from model fits for the binomial variance and configural weighting models. Mean difference between predicted and observed fallacy rates are shown in brackets. Correlations ran across all participants and all conjunction/constituent (or disjunction/constituent) pairs. All correlations were significant, but the binomial variance model had a higher correlation with observed fallacy rates. Both models tended to overestimate fallacy rates for both conjunctions and disjunctions (positive differences between predicted and observed fallacy rates) but the binomial variance model’s predicted fallacy rates were closer to the observed rate in all cases. Note that Experiment 4 did not include disjunctions, and so gives no rates for disjunction fallacy occurrence.

| Exp 2 set A | conj. fallacy | disj. fallacy |
|-------------|---------------|---------------|
|              | 0.65 (0.06)   | 0.68 (0.08)   |
| Exp 2 set B | 0.42 (0.14)   | 0.45 (0.09)   |
| Exp 3       | 0.62 (0.20)   | 0.68 (0.14)   |
| Exp 4       | 0.41 (0.14)   |               |

### 8. General discussion
The aim of the paper was to examine variability in probability estimation and its relationship to two well known cognitive biases - the conjunction and disjunction fallacies. To this end, we carried out four experiments; the first a study of variance and how different response formats effects probability judgements, the second was a study of the internal variance in probability estimation which was tested by giving the participants repeated judgement tasks. Both these experiments used description-based stimuli consistent with other description-based studies of the conjunction and disjunction fallacy. The third experiment focused on the roles of probability values and sample size on variance, estimation accuracy and fallacy rates. The final experiment again looked at the role of probability values on fallacy rates and how question type influences estimation. Experiments 3 and 4 both employed repeated judgements to understand the internal variance in participant estimates and the effect of $p$ values on that variance. Each had stimuli with observable objective probabilities so we could investigate estimate accuracy. This makes these experiments somewhat novel in research on cognitive biases. However, fallacy rates observed for both were in line with more traditional research stimuli in this field so we believe that they are appropriate.

Results showed that variability of the estimate is a key indicator of whether a fallacy response will occur. Overall, the complex statements showed higher levels of variability with statistically significant levels observed for most occasions. Approximately 70% of the complex statements were more variable than their constituent counterparts across all experiments. A small number of the simple...
statements had higher variance than the complex statements. For the description-based experiments, this occurred most frequently when the constituent had a high probability and the conjunction CI typically had no overlap with the constituent CI (e.g. P(Cloudy) vs P(Cloudy ∧ Snowy)). Statistically significant higher levels of variance are observed in some constituents with extremely low fallacy rates (0–2%).

The disjunctive statements weren’t any more variable than the conjunctive statements and very similar fallacy rates were recorded in the description experiments. No clear difference in variability can be observed between conjunction and disjunctions. Higher variance in the complex items is also observed in the visual stimuli, with both the conjunction and disjunction being more variable than the constituent. As the participants produced repeated estimates in a number of experiments, we could also analyze individual variability and its relation to fallacy rate. A consistent observation across the experiments is that participants that were more variable across their own responses for complex items were more likely to make repeated fallacy responses.

In the final experiment, we were able to compare the variance of conjunctions versus their constituents (P(A)) and value matched single events (P(C)). Here, we saw the same reported higher variance in the conjunction versus its own constituent (P(A ∧ B) v. P(A)) that we reported in the previous experiments. The variance for the P(A ∧ B) v. P(C) comparisons also found higher variance for the conjunction but for individuals, there wasn’t a significant difference between them. To date, none of the probabilistic models have

Table 16
WAIC expected log predictive density adjusted for number of parameters (elpd_WAIC) for the standard binomial variance model (constrained so that \(d_{\text{simple}} < d_{\text{complex}}\)) and the unconstrained version of that model (\(d_{\text{simple}}\) and \(d_{\text{complex}}\) take independent values) in Experiments 2, 3 and 4. The two versions of the model show no significant difference in fit for three out of 4 cases: for the first case, the standard (constrained) model gives a significantly better fit.

| Target values                  | elpd_waIC (SE) |
|-------------------------------|----------------|
| Expt 2 set A                  | 114.9 (116.4)  |
| Expt 2 set B                  | −582.7 (98.1)  |
| Expt 3 excluding single events| −1017.4 (198.2)|
| Expt 4 excluding single events| −219.1 (87.0)  |

Note: ‡: \( p < 0.05 \).

**Fig. 13.** Scatterplot of \(d_{\text{simple}}\) and \(d_{\text{complex}}\) values for the 81 participants in Experiment 2, from the binomial variance model fits. The diagonal line is the line of identity.
included an explicit model of how the variance in estimates functions. Here, we presented a simple model of variance, based on the binomial theorem, that is capable of capturing the patterns of participant responding. The Binomial variance model provides good predictions of participant variance for a given estimate and it demonstrated the importance of sample size for probability judgements, with estimates taken from larger sample sizes much less variable than estimates taken from samples whose size was small.

Conjunction fallacy rates across these experiments ranged from 0% to 68% depending on the stimulus. Similar rates of disjunction fallacies observed for the sample, ranging from 0% to 71%. These values are both in line with other research findings and the predictions of the PTN model. In experiments 1 and 2, it appeared that participants were most likely to produce a fallacy if their subjective estimates for the constituent and conjunction are close to each other, e.g. the 65% fallacy rate observed for P(Snowy) vs P(Cloudy ∧ Snowy) in experiment 2 was the highest observed for the experiment despite both sets of average estimates being low. Very low fallacy rates were likely to be observed when the constituent and conjunction estimates were unlikely to overlap. Further exploration of this trend in experiment 3 confirmed these findings. In experiment 4, we were able to control fallacy rates by manipulating the distance \((P(A ∧ B) − P(A))\) and low fallacy rates resulted. Here we see that rather than high constituent values being correlated with low fallacy rates and low constituent values being correlated with high fallacy rates, the estimate difference between the constituent and complex item is correlated with the fallacy rate.\(^6\)

Analysis of the participant estimates proved that they were internally variable. Typically, their repeated probability estimates for an item were similar, but not identical, to each other. This variability in estimates meant that participants that were often inconsistent when they produced fallacies; if they did produce a fallacy, typically they produced it for a number of occasions but not all of them. Of the possible fallacy responses (e.g. producing the fallacy between one to five times for a given conjunction or disjunction for experiment 2), a fallacy response for all of the occasions was the least likely to occur. With the larger repetitions, the likelihood of participants producing 100% fallacy rates fell. In experiments 3 and 4, no participant produced a fallacy response on all occasions. Typically, for a fallacy response to occur, one of two things should occur: the constituent and complex item estimates should be close to each other and the complex item should be more variable than its constituent. A fallacy response may occur when either of these items are present but the highest fallacy responses typically occurred when both were observed for the same estimates.

One of the most fascinating results on these experiments (particularly experiments 3 and 4) is that they revealed how good participants are at estimating probabilities. Typically, the participants produced estimates that were accurate for all of the estimation tasks presented to them. This sophistication of estimation is somewhat unexpected, particularly for research on cognitive biases, which makes a point of demonstrating the myriad of ways which humans are poor reasoners. What we find here is reasoners that are skilful, even for novel stimuli with a degree of precision that, heretofore, hasn’t been recognised in the literature. In addition to this, participant estimates are consistent with probability theory, in terms of the Addition Law expression, in all three experiments where that expression could be calculated.\(^7\) In all three, we found good compliance with the addition law, each \(A, B\) combination producing values that were close to, and varied around, the required value of 0, alongside significant conjunction and disjunction fallacy occurrence. This demonstrates that high conjunction and disjunction fallacy rates cannot be taken as evidence that people do not reason in a logical and reasonable fashion - that is, that their reasoning is always contra to probability theory. The results here demonstrate that both scenarios can occur concurrently and are not, in fact, contradictory. That probability estimates are simultaneously accurate, consistent with probability theory and produce fallacies is a major challenge to the heuristics accounts of the fallacy. Currently, noise approaches are better able to account for these results than the more traditional heuristics accounts.

9. Conclusions

The findings of this study can be taken as evidence that cognitive biases can be explained by errors in a rational probabilistic reasoning process rather than a heuristic process. Humans are good and accurate reasoners of both familiar and novel scenarios and their failings in reasoning - conjunction and disjunction fallacies in this case - arise due to a confluence of high variability in complex items and small differences in probability values. From these observations, we can conclude that probabilistic models are capable of predicting a range of biases and that they provide a coherent framework for future work on reasoning errors.

10. Financial disclosure

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare no potential conflict of interests.

Appendix A

In the main text we assume a noise rate of \(d\) for probability estimation for simple events \(P(A)\) and an increased rate of \(d + Δd\) for complex events \(P(A ∧ B)\) and \(P(A ∨ B)\). This assumption of an increase in noise \(Δd\) for complex events is very much a first

---

\(^6\) \(P(A ∧ B) − P(A)\) and \(P(B) − P(A ∨ B)\)

\(^7\) Participants didn’t provide disjunction estimates for experiment 4 so the addition law was not calculated for those estimates.
approximation. In this appendix we derive more detailed expressions for the increase in noise rate for complex events in the noisy
frequentist model.

This model assumes that people estimate the probability of some event \( A \) by randomly sampling items from memory, counting the
number that are instances of \( A \), and dividing by the sample size. The model assumes that events have some chance \( d < 0.5 \) of randomly
being read incorrectly; this random error results in an average noisy estimate for the probability of \( A \) of:

\[
(P_e(A)) = (1 - 2d)P(A) + d 
\]  
(A.1)

For conjunctive events \( A \land B \) this counting process may take place in two different ways. If the complex event is itself a familiar, already
known category, we can treat the conjunctive category as ‘integral’ and simply count items as members of that category directly, just as for
simple events. For conjunctive events \( A \land B \) that are treated in this way we get an average noisy estimate for the probability of \( A \land B \) of

\[
(P_e(A \land B \text{ integral})) = (1 - 2d)P(A \land B) + d 
\]  
(A.2)

Similarly, for disjunctive events \( A \lor B \) that represent an existing familiar category and can be treated as integral we get average noisy
probability estimate of

\[
(P_e(A \lor B \text{ integral})) = (1 - 2d)P(A \lor B) + d 
\]  
(A.3)

Note that these expressions satisfy the addition law: substituting Eqs. (A.2) and (A.3) into the Addition Law we get

\[
(P_e(A)) + (P_e(B)) - (P_e(A \land B \text{ integral})) - (P_e(A \lor B \text{ integral})) = (1 - 2d)(P(A) + P(B) - P(A \land B) - P(A \lor B)) = 0 
\]

as required in standard probability theory.

We can also, however, make this decision about membership in \( A \land B \) (or \( A \lor B \)) by treating the category as ‘separable’: by
separately checking whether the item is an instance of \( A \) (subject to noise rate \( d \)) and whether that item is an instance of \( B \) (subject to
the same noise rate \( d \)). Items which are read as \( A \) and separately read as \( B \) are labelled as instances of \( A \land B \): the probability estimate
for the conjunction is obtained by counting such labelled items and dividing by the sample size (and similarly for disjunctions).

There are at 3 possible locations of error in this ‘separable’ case: error in reading an item as \( A \), error in reading an item as \( B \), and
finally, error in reading an item as \( A \land B \). This last form of error arises when, for example, an item which was labelled as an instance of \( A \land B \) is mistakenly read as a non-instance during counting, or similarly, when an item that was not labelled as an instance of \( A \land B \) is mistakenly read as an instance during counting.

We assume, for simplicity, that all three types of error occur randomly at the same rate, \( d \). We calculate the noisy probability estimate for \( A \land B \) under these three sources of error by first giving an expression for the probability of an item being labelled as
\( A \land B \) under the first two forms of error. We then use that expression to get the average noisy estimate \( (P_e(A \land B)) \) of given random
error in reading of these labels.

We calculate the probability of a given randomly sampled item being labelled as an instance of a separable conjunction \( A \land B \) as follows. We take \( P(\text{labelled } A \land B | A \land B) \) to represent the probability of an item being labelled as \( A \land B \), given that the item truly is an
instance of \( A \land B \); we take \( P(\text{labelled } A \land B | \neg A \land B) \) to represent the probability of an item being labelled \( A \land B \), given that the item
truly is an instance of \( A \) but is not an instance of \( B \), and so on. We begin by noting that the total probability of a randomly sampled item
being labelled \( A \land B \) is obtained by summing over all possibilities for that item, each weighted by their probability of occurrence:

\[
P(\text{labelled } A \land B) = P(\text{labelled } A \land B | A \land B \land P(A \land B) + P(\text{labelled } A \land B | A \land B \land \neg B) P(A \land \neg B) + P(\text{labelled } A \land B | \neg A \land B \land P(\neg A \land B) + P(\text{labelled } A \land B | \neg A \land B \land \neg B) P(\neg A \land \neg B) 
\]

An item that is truly an instance of \( A \land B \) will be labelled \( A \land B \), in this separable process, only if both constituent events are read
correctly (with no random error); this occurs with probability \((1 - d)^2\), and we have

\[
P(\text{labelled } A \land B | A \land B) = (1 - d)^2 
\]

An item that is truly an instance of \( A \land \neg B \) (or \( \neg A \land B \)) will be labelled \( A \land B \) only if the one constituent is read correctly but the other
is read incorrectly; this occurs with probability \((1 - d)d\), and we have

\[
P(\text{labelled } A \land B | A \land \neg B) = (1 - d)d 
\]

Finally, an item that is truly an instance of \( \neg A \land \neg B \) will be labelled \( A \land B \) only if the both constituents are read incorrectly; this occurs
with probability \(d^2\), and we have

\[
P(\text{labelled } A \land B | \neg A \land \neg B) = d^2 
\]

and substituting into the overall expression above we get

\[
P(\text{labelled } A \land B) = (1 - d)^2P(\text{A } \land B) + (1 - d)d[P(\text{A } \land B) + P(\neg A \land B)] + d^2P(\neg A \land \neg B) 
\]
or simplifying

\[
P(\text{labelled } A \land B) = (1 - 2d)^2P(\text{A } \land B) + d(1 - 2d)[P(\text{A } + P(\text{B})] + d^2 
\]

Finally, with random error at the same rate \( d \) in counting instances that have been labelled \( A \land B \), we get
\[ P_e(A \land \text{Biseparable}) = (1 - 2d)P(\text{labelled } A \land B) + d \]
\[ = (1 - 2d)[(1 - 2d)^2P(A \lor B) + d(1 - 2d)[P(A) + P(B)] + d^2] + d \]  
(A.4)

A similar derivation gives a probability estimate for a separable disjunction of
\[ P_e(A \lor \text{Biseparable}) = (1 - 2d)[(1 - 2d)^2P(A \lor B) + d(1 - 2d)[P(A) + P(B)] + 2d - d^2] + d \]  
(A.5)

Note that these expressions approximately satisfy the addition law; substituting Eqs. (A.4) and (A.5) into the Addition Law we get
\[
\begin{align*}
(P_e(A)) + (P_e(B)) - (P_e(A \land \text{Biseparable})) - (P_e(A \lor \text{Biseparable})) &= (1 - 2d)[P(A) + P(B) - (1 - 2d)^2P(A \land B) + P(A \lor B)] \\
&- 2d(1 - 2d)[P(A) + P(B)] - 2d \\
&= 2d(1 - 2d)[P(A) + P(B)] - 2d(1 - 2d) \\
&= 2d(1 - 2d)(P(A) + P(B) - 1) = 0 
\end{align*}
\]  
(A.6)

and since \(-1 \leq P(A) + P(B) - 1 \leq 1\) necessarily holds, the average value for this expression across a wide range of probabilities \(P(A), P(B)\) will be 0, just as required in standard probability theory.

To combine these various measures, we need some way of estimating the probability that a given pair of events \(A\) and \(B\) will be treated separably or integrally. A natural way to estimate this probability is to say that the probability of \(A\) and \(B\) being treated separably is simply equal to the probability of those events occurring separately from each other, which we can write as
\[ P(\text{separable}) = P(A \land \neg B) + P(\neg A \land B) \]

The higher this probability the more likely it is that \(A\) and \(B\) will occur separately, and the more likely it is that \(A\) and \(B\) will be treated as separable events. Similarly, we can say that the probability of \(A\) and \(B\) being treated integrally is equal to the probability of those events occurring together (if \(A\) occurs, \(B\) occurs; if \(A\) does not occur, \(B\) does not occur; and vice versa), which we can write as
\[ P(\text{integral}) = P(A \lor B) + P(\neg A \land \neg B) = 1 - P(\text{separable}) \]

The higher this probability, the more likely it is that \(A\) and \(B\) will only ever be seen together, and so will be treated as a single integral event. Given these probabilities we get, as our overall expression for the average noisy probability estimate for a conjunction, the expression
\[ P_e(A \land B) = P_e(A \land \text{Biseparable})P(\text{separable}) + P_e(A \land \text{Bintegral})P(\text{integral}) \]  
(A.7)

and similarly
\[ P_e(A \lor B) = P_e(A \lor \text{Biseparable})P(\text{separable}) + P_e(A \lor \text{Bintegral})P(\text{integral}) \]  
(A.8)

These expressions are clearly very complex (we are not going to expand them here). One thing worth noting, however, is that these expressions again approximately satisfy the Addition Law; substituting Eqs. (A.7) and (A.8) into the Addition Law we get
\[
\begin{align*}
(P_e(A)) + (P_e(B)) - (P_e(A \lor B)) &= P(\text{separable})[P_e(A)] + (P_e(B)) - (P_e(A \land \text{Biseparable})) - (P_e(A \lor \text{Biseparable})) \\
&= 2d(1 - 2d)P(\text{separable})[P(A) + P(B) - 1] = 0 
\end{align*}
\]  
(A.9)

with the last line following from Eq. (A.6).

Given the complexity of Eqs. (A.7) and (A.8), we approximate them by assuming a noise rate of \(d\) for simple probability estimates \(P_e(A)\) and \(P_e(B)\) but an increased noise rate of \(d + \Delta d\) for conjunctive and disjunctive estimates, giving
\[
\begin{align*}
P_e(A \land B) &\approx [1 - 2(d + \Delta d)]P(A \land B) + (d + \Delta d) \\
P_e(A \lor B) &\approx [1 - 2(d + \Delta d)]P(A \lor B) + (d + \Delta d) 
\end{align*}
\]  
(A.10)(A.11)

We use this particular \(d + \Delta d\) approximation for three reasons. First, numerical simulations show that Eqs. (A.7) and (A.8) are regressive towards 0.5 in a way that is systematically stronger than that of matching single probability estimates as given in Eq. (A.1) (see Fig. 14). The use of \(\Delta d\) increase in error rate captures this increased regression. To conduct these simulations we generated 2000 randomly selected objective probability sets, all consistent with standard probability theory. Probability sets were produced by selecting random values for a triplet of objective probabilities \(P(B), P(AB)\) and \(P(\neg A \land B)\) (chosen uniformly in the range 0..1), and used these values to calculate associated probabilities in each set by applying the equations of probability theory (so that \(P(A) = P(AB)P(B) + P(\neg A \land B)(1 - P(B)), P(A \land B) = P(AB)P(B)\), and so on). Mean noisy estimates for each set was calculated by applying Eqs. (A.7) and (A.8) to the probabilities in each set (with \(d = 0.1\)). Fig. 14 shows the mean noisy estimate produced in this way, and shows that these estimates are systematically more regressive towards 0.5 than matching single probability estimates.

Second, we use this \(\Delta d\) approximation because it can produce noisy probability estimates which agree reasonably well with those produced by the more complex expressions derived above. Using the same numerical simulation data described above we found a very close fit between Equations A.10 and A.11 for \(d = 0.1\) and \(\Delta d = 0.04\) (correlation: \(r = 0.99\), root mean squared difference between values of \(RMSD = 0.018\)). Exactly the same close fit obtained between Eqs. (A.7) and (A.11) with \(d = 0.1\) and \(\Delta d = 0.04\).
\[
\text{Fig. 14. Graph of true probability } P(A \land B) \text{ versus mean noisy estimate } \langle P_k(A \land B) \rangle \text{ (left) and true probability } P(A \lor B) \text{ versus mean noisy estimate } \langle P_k(A \lor B) \rangle \text{ (right) for 2000 randomly selected probability sets. Probability sets were produced by selecting random values for a triplet of probabilities } P(B), P(A|B) \text{ and } P(A) \text{ (each having a random value chosen uniformly in the range } 0..1), \text{ and using these values to calculate associated probabilities in each set by applying the equations of probability theory (so that } P(A \land B) = P(A|B)P(B), P(A) = P(A|B)P(B) + P(A) (1 - P(B)), \text{ and so on).}
\]

Mean noisy estimates for each set were calculated by applying Eqs. (A.7) and (A.8) to the probabilities in each set (with } d = 0.1). The value of } \langle P_k(A \land B) \rangle \text{ or } \langle P_k(A \lor B) \rangle \text{ for each set is represented by a small black dot: the speckled areas represent the distribution of such values associated with each true probability } P(A \land B) \text{ (or } P(A \lor B)). \text{ White circles represent noisy estimates produced by applying the 'd' equations } (1 - 2d)P(A \land B) + d \text{ and } (1 - 2d)P(A \lor B) + d, \text{ also with } d = 0.1. \text{ The solid 45° line is the line of identity. The graph shows that mean noisy estimates have a higher degree of regression towards 0.5 (black dots falling above white circles when true probability is below 0.5, and below white circles when true probability is above 0.5). This indicates that the complex expressions in Eqs. (A.7) and (A.8) can be approximated by a simpler '}(d + \Delta d)' expression given in Eqs. (A.10) and (A.11), with the increase in error } \Delta d \text{ capturing this increase in regression.}

\[
(r = 0.99, \text{RMSD} = 0.018). \text{ Similar fits hold for different values of } d.
\]

Finally, we use this particular } \Delta d \text{ approximation because it makes predictions about the addition law identity that match those from Eqs. (A.7) and (A.8). Substituting Eqs. (A.10) and (A.11) into the Addition Law expression we get}

\[
\langle P_k(A) \rangle + \langle P_k(B) \rangle - \langle P_k(A \land B) \rangle - \langle P_k(A \lor B) \rangle
\]

\[
= 2\Delta d[P(A) + P(B) - 1] - 0
\]

and this approximation gives a predicted value for the Addition Law whose form follows that given in Eq. (A.9).

References

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. Journal of Personality and Social Psychology, 65, 1119.

Bearden, N. J., & Wallsten, T. S. (2004). Minerva-DM and subadditive frequency judgments. Journal of Behavioral Decision Making, 17, 349–363.

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. Mind & Language, 19, 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, 157–171.

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

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76.

Costello, F., & Mathison, T. (2014). On fallacies and normative reasoning: When people’s judgements follow probability theory. Proceedings of the 36th annual meeting of the Cognitive Science Society (pp. 361–366).

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

Costello, F., & Watts, P. (2016). People’s conditional probability judgments follow probability theory (plus noise). Cognitive Psychology, 89, 106–133.

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. Journal of Behavioral Decision Making, 30, 304–321.

Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. Cognitive Psychology, 100, 1–16.

Costello, F., & Watts, P. (2019). The rationality of illusory correlation. Psychological Review, 126, 437.

Costello, F., Watts, P., & Fisher, C. (2018). Surprising rationality in probability judgment: Assessing two competing models. Cognition, 170, 280–297.

Dawson, N. V., & Arkes, H. R. (1987). Systematic errors in medical decision making. Journal of General Internal Medicine, 2, 183–187.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). Minerva-DM: A memory processes model for judgments of likelihood. Psychological Review, 106, 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, 519.

Eva, K. W., & Norman, G. R. (2005). Heuristics and biases: Biased perspective on clinical reasoning.

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

Eva, K. W., & Norman, G. R. (2005). Heuristics and biases: Biased perspective on clinical reasoning.

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

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. Psychological Research, 50, 123–129.
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, 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, 1–20.

Galástel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*, 121, 96.

Gavranski, L., & Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. *Journal of Personality and Social Psychology*, 61, 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, 275–305.

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

Johnson, D. D., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*, 28, 474–481.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, 93, 1449–1475.

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

Korobkin, R., & Ulen, T. (2000). Law and behavioral science: Removing the rationality assumption from law and economics. *California Law Review*, 88, 1051.

Marchiori, D., Di Guida, S., & Erev, I. (2015). Noisy retrieval models of over-and undersensitivity to rare events. *Decision*, 2, 82.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 93, 1449–1475.

Nilsson, H., Juslin, P., & Winman, A. (2016). Heuristics can produce Surprisingly Rational Probability Estimates: A commentary on Costello and Watts (2014). *Psychological Review*, 123(1), 103–111.

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, 517.

Oliver, A. (2013). From nudging to budging: Using behavioural economics to inform public sector policy. *Journal of Social Policy*, 42, 685–700.

Reeves, T., & Lockhart, R. S. (1993). Distributional versus singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General*, 122, 207.

Rieskamp, J., & Otto, P. E. (2006). SLE: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, 135, 207.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E. J. (2013). Testing adaptive toolbox models: A bayesian hierarchical approach. *Psychological Review*, 120, 39.

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

Söllner, A., Bröder, A., Glückner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica*, 146, 84–96.

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

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

Teigen, K. H., Martinussen, M., & Lund, T. (1996). Linda versus world cup: Conjunctive probabilities in three-event fictional and real-life predictions. *Journal of Behavioral Decision Making*, 9, 77–93.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.

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

Vehtari, A., Gelman, A., & Gabry, J. (2018). Loo: Efficient leave-one-out cross-validation and waic for bayesian models [Computer Software]. R package, version 2.1.3, https://mc-stan.org/loo.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27, 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.

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