Strange but true: Corroboration and base-rate neglect.

Toby D. Pilditch[1,2,*], Sandra Lagator[1] and David Lagnado[1]

[1]*Department of Experimental Psychology, University College London, WC1H 0AP, UK*

[2]*University of Oxford, School of Geography and the Environment, South Parks Road, Oxford, OX1 3QY, UK*

*Corresponding Author: Correspondence should be addressed to Toby D. Pilditch, 26 Bedford Way, London, WC1H 0AP, UK. Electronic mail can be sent to t.pilditch@ucl.ac.uk.

Declarations of interest: None.

Abstract

How do we deal with unlikely witness testimonies? Whether in legal or everyday reasoning, corroborative evidence is generally considered a strong marker of support for the reported hypothesis. However, questions remain regarding how the prior probability, or base rate, of that hypothesis interacts with corroboration. Using a Bayesian Network model, we illustrate an *inverse* relationship between the base rate of a hypothesis, and the support provided by corroboration. More precisely, as the base rate of hypothesis becomes more unlikely (and thus there is lower expectation of corroborating testimony), each piece of confirming testimony provides a non-linear increase in support, relative to a more commonplace hypothesis – assuming independence between witnesses. We show across three experiments that lay reasoners consistently fail to account for this impact of (rare) base rates in both diagnostic and intercausal reasoning, resulting in substantial *underestimation* in belief updating. We consider this a novel demonstration of an inverted form of base rate neglect. We highlight the implications of this work for any scenario in which one cannot assume the confirmation or disconfirmation of a reported hypothesis is uniform.

Keywords: base rate neglect, coherence, cognitive bias, evidential reasoning, probabilistic reasoning

1. Introduction

Consider the following. You are investigating a possible break-in at a home in the city centre, where the front door has been knocked down. At present, you are trying to determine a description of the possible suspect. Several homes on the opposite side of the street overlook the scene. In each of the three overlooking houses, there is a witness. All three witnesses claim to have been drawn to the window overlooking the street by the noise of the front door being knocked down, and have since stayed within their homes (as such, we may consider them independent of each other, for now). You interview each witness in turn. Now, consider two possible cases:

1) Each witness reports that the suspect was a male wearing a black hoodie.
2) Each witness reports that the suspect was a male in a clown outfit.

In case 1), we may consider the description of the suspect a fairly likely account (i.e. one could easily imagine a potential burglar being a male, dressed in such a manner). Conversely, in case 2), the possibility that the suspect was a male dressed in a clown outfit seems far less likely.

At the point of only hearing from one witness in each case, we are inclined to dismiss the witness in case 2) as mistaken or unreliable, whilst the witness in case 1) lends some support to a likely scenario. However, corroboration from a second witness has a different impact on each case. In case 1) the second witness adds some further support to an already likely scenario, but in case 2) we *again* receive a seemingly unlikely report of a man dressed as a clown. Crucially, given that the first and second witnesses have observed (and are reporting) independently of one another, it now becomes substantially more probable that the suspect was in fact dressed as a clown. This can be reasoned via the fact that it is substantially more likely that the two witnesses are reporting a clown outfit *because there really was a suspect in a clown outfit* than if both witnesses are unreliable and coincidentally made up exactly the same *unlikely story*. Of additional importance, we should therefore substantially increase a) our estimation that these witnesses are in fact reliable, and b) that the third witness is in fact *likely* to report the suspect as being dressed in a clown outfit.

Let us now revisit case 1) with this intuition in hand. Here, there remains a reasonable probability that the two witness reports gathered so far could be due to both witnesses making up the same story (i.e. if one were to draw a potential burglary suspect from one's imagination, one of the most probable descriptions would be a male in a black hoodie). Given this, the degree to which we update the hypothesis of the suspect being a male in a hoodie is not as substantial as in case 2), and in tandem, we cannot infer so strongly that the two witnesses are reliable, nor that the third witness will corroborate their story.

The above example brings together the elements of base rates (how likely are the reported hypotheses) and corroboration among witnesses. This has been addressed in philosophy and legal reasoning under the topic of coherence, albeit typically focused on the likelihood of a conjunction of components within a testimony (Cohen, 1977; Olsson, 2002, 2005; Bovens & Hartmann, 2003, 2005, 2006; Douven & Meijs, 2007). Critically, this effect is based on the fact that as a (reported) hypothesis becomes more *unlikely*, corroboration among independent witness becomes *more effective* in providing support to that hypothesis. For example, if looking for the body of a murder victim in Tokyo, we would be more surprised by two witnesses independently indicating the same specific house (and raise our confidence in the body location to a greater degree as a consequence), rather than the same broad area of the city (Bovens & Hartmann, 2003; see Harris & Hahn, 2009 for an empirical demonstration of lay reasoners' fidelity to these coherence predictions). We now turn to outline each of these elements (base rates and corroboration among witnesses) as phenomena within the study of evidential and probabilistic reasoning, before illustrating empirically the way in which lay reasoners deal with their conjunction.

1.1. Base Rate Neglect

In belief updating, there are two critical components to accurately updating the probability of a hypothesis, given new evidence (i.e. the posterior probability, described in Bayesian terms as P(H|E)). First, the “strength” of that evidence, or more precisely, the conditional probability of observing that evidence, given the hypothesis is true, P(E|H), divided by the probability of that evidence occurring when the hypothesis is false, P(E|¬H), known as the likelihood ratio (LR). For example, an LR of 2 dictates the evidence is twice as likely to be due to the hypothesis than not, whilst an LR of 1 means the evidence is equally likely whether the hypothesis is true or not. Thus, the former increases the probability of the hypothesis being true (given the evidence is observed), whilst it remains unchanged in the latter. Second, one also needs to specify the probability of the hypothesis *prior to observing the evidence*, known as the prior probability – or base rate – of the hypothesis, P(H). In Bayes theorem – a formalism for optimal inference under uncertainty – the odds form of these two components (prior and LR) are multiplied to reach an accurate posterior probability (Pearl, 1988).

Research in psychology has shown that in performing belief updating, people often fail to account for this prior probability (Kahneman & Tversky, 1973; Tversky & Kahneman, 1981), known as base rate neglect. For example, in medical diagnosis, reasoners often overestimate the probability of a patient having a disease given a positive test result, by focusing on the probability of a positive test result given the disease, and failing to appropriately account for the rarity of the disease (Gigerenzer et al., 2007). Another classic example is the taxi cab problem:

"A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

(a) 85% of the cabs in the city are Green and 15% are Blue.

(b) a Witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time" (Tversky & Kahneman, 1981, pp. 9-10).

Studies have shown that participants are more likely to give a response that approximates reliability of the witness (around 80%), and very few provide the Bayesian solution that accounts for the base rate information (around 40%; Kahneman & Tversky, 1973; Bar-Hillel, 1980; Lyon & Slovic, 1976). In our study, one of the scenarios has a very low prior likelihood (~3%), while the other has a higher prior likelihood (50%). The impact of the base rates in these scenarios is such that, after corroboration, the scenario with a lower base rate becomes more likely than the scenario with a higher base rate. If participants were to neglect a low base rate, as was observed in prior studies, we would expect them to underestimate the impact of corroboration in the low-base-rate scenario more so than in the high base rate scenario.

Crucially, whether demonstrated in medical diagnosis, investigative and forensic inference, or more abstract probabilistic reasoning tasks, this effect is typified by the failure to account for a *rare or unlikely prior probability*, and thus the posterior probability is *overestimated*. Placing base rate neglect in conjunction with Cohen's (1977) supposition (supported by the Bayesian normative framework) that the coincidence of multiple independent reports corroborating something unlikely should boost confidence over and above equivalent corroboration of something more likely, we would expect the following: If participants were to neglect a low base rate, as was observed in prior studies, we would expect them to underestimate the impact of corroboration in the low-base-rate scenario more so than in the high base rate scenario. Or more succinctly, in the case of corroboration, we expect a counterpole (*under*estimation) form of base rate neglect.

1.2. Corroboration and Witness Testimony

Coherence has been explored using a number of approaches (see e.g. Shogenji, 1999; Fitelson, 2003; and Glass, 2002; 2007). For example, Holyoak and Simon (1999) and Simon, Stenstrom and Read (2015) have investigated coherence within the context of constraint satisfaction models of reasoning (McClelland & Rumelhart, 1981). These connectionist models suggest that reasoning resembles a network of possible beliefs, all linked through excitatory and inhibitory links. The important aspect of these models is that the spread of activation is bidirectional rather than unidirectional. Holyoak and Simon investigated whether such bidirectional activation can be observed empirically. In their study, participants were presented with a number of arguments, and asked to judge the extent to which they believe an argument supports a particular conclusion. In the next stages, participants were asked to read a particular case, come to a verdict, and make judgments about the arguments for the second time. The results showed that, after reaching the verdict, the judgments of arguments shifted towards a particular conclusion. The findings were interpreted as evidence of both forward (arguments to conclusions) and backward (conclusions to arguments) spreads of activation.

In the present work we focus on the Bayesian Network framework (BNs). This formalism uses directed graphs to represent probabilistic dependencies between hypotheses and evidence, and thus can capture complex interrelations between hypotheses and multiple witness testimonies (Pearl, 1988; Fenton & Neil, 2012). BNs can also represent the reliability of evidential reports by introducing exogenous variables (Bovens & Hartmann, 2003; Neil, Fenton, & Nielson, 2000; Lagnado, Fenton, & Neil, 2013). The BN framework uses Bayesian updating to determine the impact of corroboration between witnesses (of which, we may consider "coherence" effects a subset) on both diagnostic and predictive reasoning. To illustrate, Figure 1 below depicts a BN that captures three conditionally independent witnesses (Rep_{1-3}) reporting on hypothesis H, each with their own reliability (Rel_{1-3}).[1]

.

[1] More precisely, each reporting witness (Rep_{1-3}) is conditionally independent of each other given the hypothesis H (a necessary component of coherence effects, but see Bovens and Hartmann, 2003, pp 64 for the influence of relaxing this assumption, and Madsen, Hahn, & Pilditch, 2018 for an empirical demonstration).

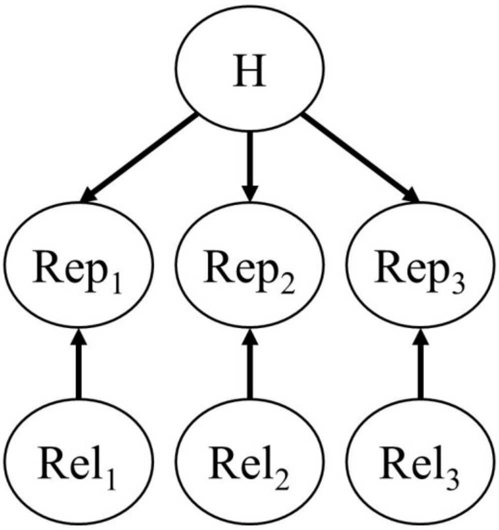

Fig. 1.

Graphical structure of three (conditionally) independent witnesses (Rep_{1-3}), each with their own reliability (Rel_{1-3}), informing on a hypothesis (H).

Each witness report is thus a “common effect” of the hypothesis H and the reliability of the witness Rel. This causal structure has been used in legal reasoning as a “reliability idiom” (Fenton, Neil & Lagnado, 2013) to model the impact of witness testimony.

1.3. Formalism and Present Research

In addition to the graph, a BN has a conditional probability table (CPT) for each variable in the model. In the case of the reliability and hypothesis components, these are just prior probabilities (both starting at 50% in the example below). For reports (i.e. testimony provided by the witness), the conditional probability table (CPT) in Table 1 uses the break-in example with the hoodie case as an example for how the two parents H and Rel_i probabilistically influence the effect variable Rep_i.

Table 1.

Conditional probability table (CPT) for Witness 1, given a hypothesis (P(H = Hoodie) = 0.5) and an exogenous reliability (Rel), both of which may be either true or false.

	Rel$_1$ = True		**Rel$_1$ = False**	
	H = ¬Hoodie	H = Hoodie	H = ¬Hoodie	H = Hoodie
Rep$_1$ = Hoodie	0	1	.5	.5
Rep$_1$ = ¬Hoodie	1	0	.5	.5

In this example, we assume that if the witness is reliable (Rel$_1$ = True), then they correctly report whether H is true or false (Rep$_1$ = Hoodie, when H = Hoodie, and Rep$_1$ = ¬Hoodie, when H = ¬Hoodie). Conversely, when the witness is not reliable (Rel$_1$ = False), the witness essentially reports the hypothesis as true or false *at random*. This particular set of parameters is an example of a witness that – when paying attention – will provide a correct report with perfect honesty, but when not paying attention, is still obliged to provide a report, so reports based on chance. Importantly, the randomly generated report may be considered to occur with the same probability as the prior probability of that hypothesis (i.e. if it is commonplace that break-ins are perpetrated by people wearing hoodies, then it is assumed to be equivalently as likely an unreliable witness will make up a description of a hoodie being worn).

Crucially, the above example uses a hypothesis with a prior of 50% (P(H = Hoodie) = 0.5). To illustrate the influence of base rates, we should consider an improbable hypothesis (e.g. P(H = Clown) = 0.03). Turning to the CPT illustrated in Table 2 below, when the witness is reliable, (Rel$_1$ = True, left-hand pair of columns) the probability of reporting the truth does not change. However, when the witness is unreliable (Rel$_1$ = False, right-hand pair of columns), the probability of reporting *that particular hypothesis (i.e. a clown outfit)*, vs *anything other than that particular hypothesis* is not 50/50.

Table 2.

Conditional probability table (CPT) for Witness 1, given a hypothesis (P(H = Clown) = 0.03) and an exogenous reliability (Rel), both of which may be either true or false.

	Rel$_1$ = True		**Rel$_1$ = False**	
	H = ¬Clown	H = Clown	H = ¬Clown	H = Clown

Rep_1 = Clown	0	1	.03	.03
Rep_1 = ¬Clown	1	0	.97	.97

In fact, given the rarity of the hypothesis itself, it may be fairer to assume the probability of stating that hypothesis by chance is in fact 3% (i.e. $P(\text{Rep}_1|\neg\text{Rel}_1, H = \text{Clown}) = P(\text{Rep}_1|\neg\text{Rel}_1, H = \neg\text{Clown}) = .03$), and therefore reporting something *other than this particular hypothesis* is 100 – 3 = 97%. Thus, we capture the inference that it is unlikely for someone to claim that a clown committed the burglary when generated by an inattentive witness.

2. Experiment 1

To test the above supposition empirically, we used the burglary example. By providing participants with two hypothetical cases (witnesses independently reporting a culprit wearing a black hoodie / witnesses independently reporting a culprit wearing a clown outfit) and asking a qualitative comparison question (which case are you more convinced by?) repeatedly as more corroboration occurs, we could assess sensitivity to the above reasoning intuition. More precisely, whilst the black hoodie should start off being the more likely description, as corroboration among independent witnesses increases, the clown outfit should become more likely (as coincidental independent reports of a clown outfit being made up by witnesses becomes increasingly (and severely) unlikely) – we predict participants will fail to intuit this dynamic, as it relies on an accurate incorporation of the base rate of an unlikely support. We do however note that such a prediction is at this point exploratory in nature, though in line with the theoretical predictions of Cohen (1977).

2.1 Method

Participants. 60 participants were recruited and participated online through MTurk (https://www.mturk.com/). Those eligible for participation had a 95% and above approval rating from over 100 prior HITs. Participants were English speakers, located in the United States. The mean age was 33.63 (*SD* = 10.51), and 19 participants identified as female. Participants gave informed consent and were paid $0.80 for their time (*Median* = 2.85 minutes, *SD* = 4.13). All experiments received ethical approval from the UCL Research Ethics Committee [EP/2017/005].

Materials and Procedure. Participants were presented with a burglary scenario in which they must work out a description of the culprit using witness testimonies from people in the houses opposite the crime scene. Critically, participants were told that the three witnesses were all drawn to their windows overlooking the street by the noise of the break in, and remained in their homes after the event. In this way, it was highlighted that witnesses could be considered independent, and should be considered equally reliable. This reliability was operationalized as follows: Given the burglary took place at night, each witness had a 50% probability of not having actually seen anything (i.e. $P(Rel_{1\text{-}3}) = 0.5$). If a witness was reliable (i.e. had actually seen the culprit), then they would definitely report the truth (i.e. provide a correct description). If a witness was not reliable (i.e. had not actually seen the culprit), then the witness would (in wanting to be helpful) provide a made-up description. In this way, the probability of making-up a particular description is equal to the probability of that description being true *a priori* (i.e. $P(Rep{=}True|\neg Rel) = P(H)$).

Participants were then instructed to consider two hypothetical cases for the burglary scenario, one in which the culprit wore a black hoodie (A), the other in which the culprit wore a clown outfit (B). Prior to receiving any witness reports, participants then provided their estimates of the prior probability of each description being true (*What is the probability of a culprit wearing a [black hoodie / clown outfit]?*; slider from 0-100% for each). These prior probabilities could then be used to fit individual Bayesian Network models to each participant (see section 2.2. below).

Participants were then asked to make the following comparison judgment at this baseline stage, and then again as each witness provided a report (T1, T2, and T3):

Given everything you know so far, in which hypothetical account is the description of the culprit more likely to be true? [A (Hoodie) / B (Clown) / There is no difference]

Along with a confidence estimate for that judgment:

*How **confident** are you that **your response is correct**?* [Slider, 0-100%]

Importantly, when a new witness report was observed, the two hypothetical cases were displayed simultaneously (i.e. Witness 1 in Case A reported the culprit was wearing a black hoodie, and Witness 1 in Case B reported the culprit was wearing a clown outfit). This facilitated the capacity to make comparative judgments of description likelihoods. All witness reports confirmed the description of that given case. For a full description of the methods, see Supplementary Materials.

2.2 Results

Unless indicated otherwise, all experiment analyses were Bayesian and performed using the JASP statistical software (JASP Team, 2018)[2]. Taking the elicited prior probabilities for the hypothesis (burglar wore a black hoodie/clown outfit) from each participant, individually fitted Bayesian Networks (hereafter termed Behaviorally Informed Bayesian Networks; BIBNs) were created using the gRain package in R (Højsgaard, 2012). These fitted models were created because participants may bring their own adjustments to the priors stated in the materials, and these priors can influence the subsequent model predictions. The remaining structure and parameters were taken from the background information presented to all participants. These models were then used to generate normative predictions (case preferences) for each participant across elicitation stages, to be used in subsequent comparison analyses.

[2] All Bayesian analyses use a uniform prior. Here, we use the inverse Bayes Factor (BF_{10}; hypothesis over null). $BF_{10} > 3$ are considered significant, and >100 are considered "decisive" (Jarosz & Wiley, 2014). Lastly, in using BFs we may infer evidence for the null hypothesis, wherein $BF_{10} < 1/3$ is considered substantial support for the null (Dienes, 2014).

2.2.1. Case Preference Data

Using Bayesian binomial tests (*N* = 60), we first compared participant preferences (grey bars, Fig. 2) across the four stages to chance (test value = 0.33). This revealed that participants decisively preferred the hoodie case across baseline (*Prop* = 0.75), $BF_{10} > 10000$, δ = 0.745 (95% CI: [0.627, 0.842]), T1 (*Prop* = 0.65), $BF_{10} > 10000$, δ = 0.647 (95% CI: [0.523, 0.758]), and strongly at T2 (*Prop* = 0.533), $BF_{10} = 29.94$, δ = 0.533 (95% CI: [0.408, 0.654]), but no different from chance at T3 (*Prop* = 0.417), $BF_{10} = 0.421$, δ = 0.418 (95% CI: [0.3, 0.543]). In tandem, participants decisively eschewed the clown case across baseline, (*Prop* = 0.067), $BF_{10} > 10000$, δ = 0.076 (95% CI: [0.027, 0.159]), T1 (*Prop* = 0.05), $BF_{10} > 10000$, δ = 0.06 (95% CI: [0.018, 0.137]), T2 (*Prop* = 0.05), $BF_{10} > 10000$, δ = 0.06 (95% CI: [0.018, 0.137]), *and T3* (*Prop* = 0.083), $BF_{10} = 2822.75$, δ = 0.092 (95% CI: [0.037, 0.181]).

In summary, we observe a strong preference for the hoodie case, which gradually decreases across stages. Critically, this decrease does not correspond to an increase in clown case preferences at any stage.

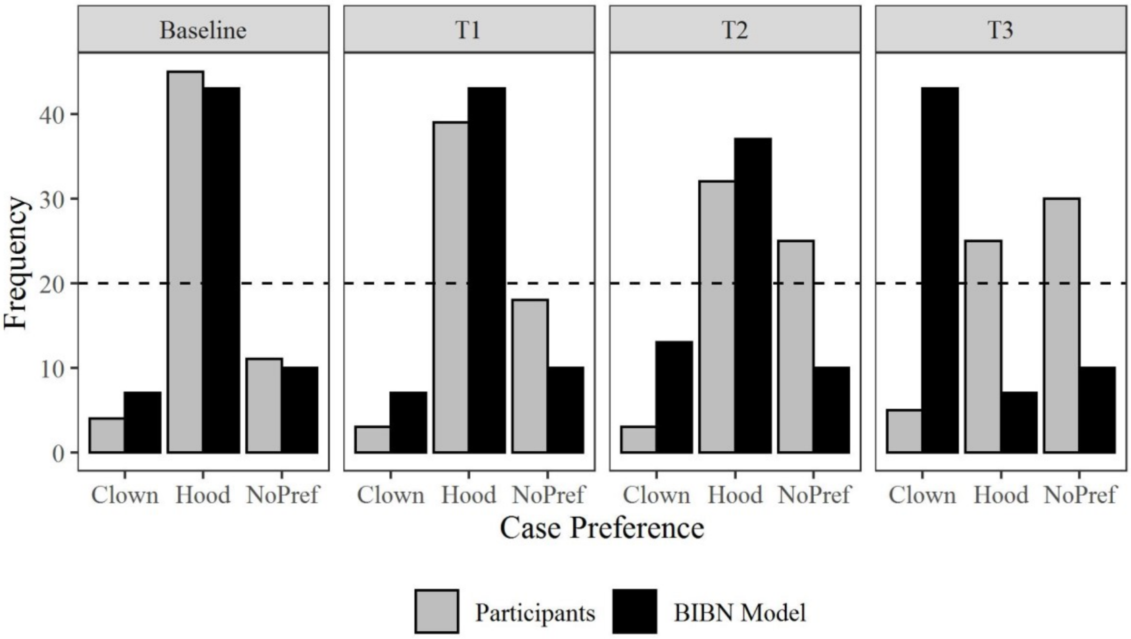

Fig. 2.

Case Preferences across evidence stages, grey bars represent participant data, black bars represent BIBN model predictions (fitted on individual level).

To determine the degree to which individual participant preferences then corresponded to their BIBN model predictions, a binary variable was created for whether a participant case preference matched the preference of their model (1 = correct; 0 = incorrect). These variables were then compared to chance (test value = 0.33) using Bayesian binomial tests across the four stages. In line with the gradually increasing disparity between participants (grey bars, Fig. 2) and their BIBN models (black bars, Fig. 2) the correctness of participant preferences are decisively above chance at baseline (*Prop* = 0.817), $BF_{10} > 10000$, $\delta = 0.81$ (95% CI: [0.7, 0.894]), T1 (*Prop* = 0.783), $BF_{10} > 10000$, $\delta = 0.777$ (95% CI: [0.663, 0.868]), and strongly at T2 (*Prop* = 0.55), $BF_{10} = 71.65$, $\delta = 0.549$ (95% CI: [0.424, 0.669]), but no different from chance at T3 (*Prop* = 0.25), $BF_{10} = 0.345$, $\delta = 0.255$ (95% CI: [0.158, 0.373]). In summary, the increased error rate at T3 corresponds to the substantially insufficient preferences for the clown case among participants, as compared to their BIBN model predictions.

2.2.2. Confidence in Case Preferences

In corroboration of the increasing error rates in judgments, using a Bayesian repeated measures ANOVA it was found that confidence in preferences decreased across stages, $BF_{10} = 140.16$. Table 3 below illustrates this gradually decreasing trend, although it should be noted that confidence in preferences generally remains high.

Table 3.

Confidence estimates across elicitation stages (*N* = 60).

Stage	*Mean*	*SD*
Baseline	81.68	19.08
T1	77.45	21.49
T2	71.68	23.33
T3	71.43	24.06

2.3 Conclusion

Using the example burglary case, we firstly demonstrate how lower base rate hypotheses can yield more support (than an a priori more likely hypothesis) as corroboration increases – via participant fitted Bayesian models. Secondly, we show participants fail to appreciate this effect using qualitative comparison preferences, and eschew the rare hypothesis despite the strong degree of independent corroboration. To complete this picture, we also find that although confidence in preferences is high across the task, the higher error rates at later stages (e.g. T3) - due to the influence of base rates and corroboration – are accompanied by a drop in participant confidence.

3. Experiment 2

Building from Experiment 1, we seek to strengthen these findings by incorporating further relevant components to the phenomena at hand. More precisely, we enrich the empirical work via the inclusion of probabilistic estimations, expanding our scope of enquiry to cover estimations of the reliability of sources, the accuracy of predictive inferences, and the influence of contradiction among sources.

To explain this in more detail we use a second set of experimental materials; two cases of a “semi-blind” roulette scenario. In the first case, the reasoner wants to know whether or not the ball lands on red or black (P(H)= .5); in the second case, the reasoner wants to know whether or not the ball lands on a specific number (e.g. the number 23, out of 36 possible numbers, P(H) ≈ .03). Having not observed the outcome directly yourself (the “blind” component), you must rely on three independent observers, each of whom have either accurately observed the outcome (reliable) or have not been paying attention (unreliable). But you don’t know which observers are reliable or not. To outline these inferences, Fig. 3 below follows the two comparison cases, common (0.5) prior (“Ball landed on red”, left-hand column), and rare (0.028) prior (“Ball landed on 23”, right-hand column) over a series of time-points as the witnesses provide their reports in sequence.

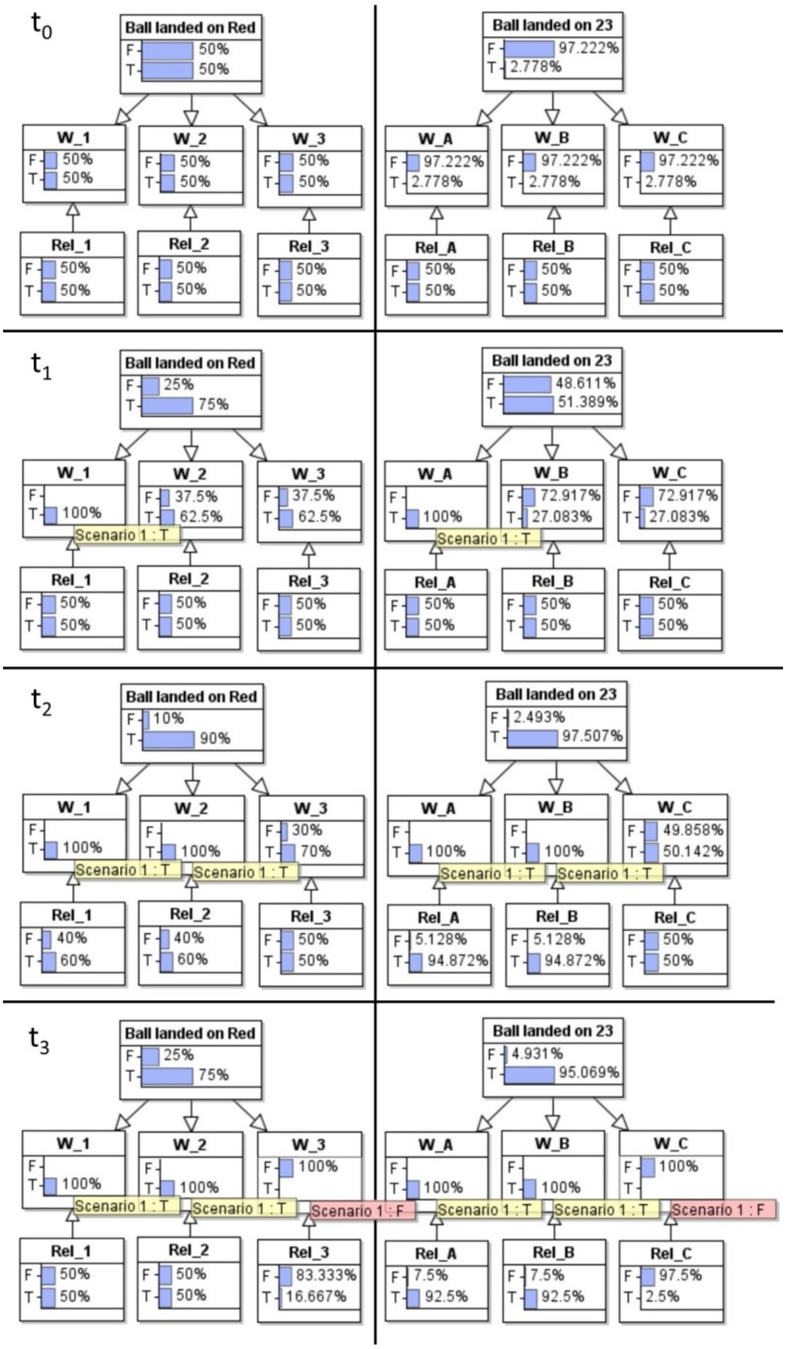

Fig. 3.

Bayesian Networks showing common (left-hand column) vs rare (right-hand column) reported hypotheses ("Ball landed on X"), reporting witnesses (W_1-3 / W_A-C) and their reliabilities (Rel_1-3 / Rel_A-C). Row t_0 reflects the baseline for each case, with no reports observed, row t_1 shows the influence of a single reporting (confirming) witness, t_2 a corroborating report from a second independent witness, and t_3 an additional contradicting witness. All reliability nodes are identical, and witness CPTs are set up as outlined in Table 1.[3]

As can be seen at t_0, before any reports come in, the probability of a witness reporting "red" is far higher than reporting the specific "number 23" – a reflection of their baseline probabilities. This reflects the surprising nature of an unlikely report.

When only a single witness has reported (t_1), we can already note that the probability of the "ball is 23" hypothesis has increased substantially (2.77% to 51.39%) relative to the more probable "ball is red" hypothesis (50% to 75%). Crucially, at this stage it is important to note that a) subsequent reports of the number 23 are still improbable, relative to "red" counterparts, and b) nothing may be inferred in regard to the reliability of the reporting witness (it remains at 50%) – to update this would be an error akin to the double-counting of evidence in updating within evidential reasoning (Schum & Martin, 1982).

Only at t_2, when a second witness has reported, may we make an inference in regard to the reliability of reporting witnesses. Now, given the difference in how (un)likely corroborating reports are a product of independent unreliability (and therefore have simply occurred by chance), the coherence between the "number 23" reports results in not only greater confidence in the respective hypotheses (97.5% for "ball is 23" vs 90% for "ball is red"), but also greater confidence that reporting witnesses are reliable (94.87% for "ball is 23" reporters, vs 60% for "ball is red" reporters).

Lastly, at t_3, a final, third witness contradicts the two previous witnesses. When comparing the rare (right-hand) to common (left-hand) base rate scenarios, there are several important effects. Firstly, the stronger coherence in the rare base rate (number) case results in a retention of the confidence in the hypothesis, whilst the contradicting witness has their reliability more severely penalized as a consequence. Reasoning intercausally, we can follow the inference

[3] Figure created using the AgenaRisk software (Agena, 2018).

that if the hypothesis is probably true (as a product of other evidence), then the contradicting report must be explained away as a product of unreliability. Secondly, the corroborating witnesses in the rare base rate scenario retain much of their reliability (again due to the higher coherence), whilst the reliability of the corroborating witnesses in the common base rate scenarios are in effect returned to the maximum uncertainty (50/50) surrounding their reliability, given the presence of a contradicting minority.

In the present work, we explicitly test lay reasoners capacity to intuit these inferences relating to base rates and corroboration. Given sequentially presented evidence, we look at diagnostic reasoning of the reported hypothesis and the reliability of the reporting witness (as further witnesses either corroborate or contradict them), and lastly the predictive reasoning of how likely a further witness is to provide a corroborative report. We look at these probability estimates within-subjects across the two cases of the “semi-blind” roulette scenario, one in which the reasoner is interested in a hypothesis of the ball landing on red or black (P(H)= .5), and the other on a specific number (e.g. the number 34, out of 36 possible numbers, P(H) ≈ .03).

We predict three hypotheses:

1. Reasoners may erroneously “double count” a single reporting witness (see e.g. Schum & Martin, 1982), such that not only is the probability of the hypothesis updated, but the probability of the witness being reliable is also (falsely) updated.[4] We note this hypothesis as explorative, given the absence of empirical data to date on this form of double counting.
2. There will be an overall underestimation in belief updating, across all sequential evidence presentations (irrespective of hypothesis base rates), in line with previous research on conservative updating (Phillips & Edwards, 1966).

[4] To the authors’ knowledge, no empirical evidence of this form of double-counting exists. However, Schum and Martin (1982) have shown that evidence which should be redundant in regards to an explanation may still be erroneously judged as probative. For this prediction, we consider (un)reliability a possible explanation of a report.

3. Critically, and key to the present work, we predict that reasoners will substantially *underestimate* the posterior probabilities, given corroborative reports, of a) the reported hypothesis, b) the reliability of corroborating witnesses, and c) the likelihood of further corroborating reports when *base rates are lower* (number case) rather than higher (color case) – i.e. an inversion of standard base rate neglect. More precisely, in standard base-rate neglect problems, the neglect of base-rates leads to people over-estimating the impact of diagnostic evidence, whereas in the current case the neglect of base rates leads to an under-estimation. We predict this based on the findings of, and rationale behind, Experiment 1 (i.e. 3a is effectively a conceptual replication attempt), in conjunction with the implication of this neglect effect to reliability (3b) and predicted testimony (3c) illustrated by the Bayesian Network model (see Fig. 3).

3.1. Method

Participants. 122 participants were recruited and participated online through MTurk (https://www.mturk.com/). Those eligible for participation had a 95% and above approval rating from over 100 prior HITs. Participants were English speakers, located in the United States. Of the 120 participants remaining, 52 were female. The mean age was 36.58 (*SD* = 12.16). Participants gave informed consent and were paid $1 for their time (*Median* = 5.07 minutes, *SD* = 4.11).

Materials and Procedure. Participants were presented with two roulette scenarios. In the color scenario, a bet was placed that the ball will land on red (P(H) = .5); in the number scenario, a bet was placed that the ball will land on number 34 (P(H) ~ 0.27). Participants were told that they cannot observe the spin outcomes themselves and should rely on reports given by three witnesses. The witnesses provided reports independently of each other and were reliable 50% of the time. Participants were told that reliability was operationalised as veracity, i.e. when a witness has paid attention (reliable), they will always report the outcome honestly, whilst an unreliable witness has not paid attention, and so reports an outcome *randomly*. All participants were presented with both scenarios, the order of which was counterbalanced between participants. In both scenarios, the three reports were presented one at a time. The first two reports supported the hypothesis (*The ball has landed on red/34*), and thus corroborated each other, while the third one contradicted the former two (*The ball has not landed on red/The ball has landed on another number (not 34)*). Participants were asked to make judgments at each of the following stages: Baseline, First report, Second report, Third report.

At Baseline, prior to seeing any reports, participants were asked to provide two estimates: 1) the probability that the ball has landed on red/34, 2) the probability of each witness being reliable. Subsequently, participants were asked to update their beliefs based on the evidence received. After each of the three reports, the following two estimates were required:

Probability of the Hypothesis: *What is the probability that the ball has landed on red/34, given everything you know so far?*

Reliability of the First Witness: *What is the probability that witness 1/A is reliable (i.e. paid attention to this outcome); given everything you know so far?*

After the first two reports, there was also a third question that required participants to make a Third Witness Prediction:

What is the probability that witness 3/C will also report red/34?

After the third report, the third question required participants to make an estimate of Reliability of the Third Witness:

What is the probability that witness 3/C is reliable (i.e. paid attention to this outcome); given everything you know so far?

At the end of each question, participants were reminded that the estimate being asked for refers to the particular instance described (based on the evidence received), and not the general likelihood (e.g., the probability of the ball landing on red in general). For a full description of the methods, see Supplementary Materials.

3.2. Results

As in Experiment 1, all analyses were Bayesian, and elicited prior probabilities for the hypothesis (ball landed on red/34) as well as witness reliability from each participant, were used to fit individual Bayesian Networks (BIBNs). As in Experiment 1, these fitted models accommodate for a) participants bringing their own adjustments to the priors stated in the materials, and b) the influence of said assumptions on subsequent model predictions. The remaining structure and parameters were taken from the background information presented to all participants. The models were then used to generate normative predictions (probability estimates) for each participant across elicitation stages, to be used in subsequent comparison analyses.

3.2.1. Probability of the hypothesis

There are three interrelated hypotheses regarding participant estimates of the hypothesis (ball was red / ball was no. 34) across the experiment: Firstly, are participants, in line with previous literature, generally conservative in their updating, given new information? Secondly, does this insensitivity occur not only when reports corroborate, but also when a report contradicts? Finally, is this insufficient updating exacerbated in cases of lower base rates (i.e. the impact of coherence)?

To test the first of these hypotheses, a Bayesian repeated measures ANOVA was performed which included all relevant factors; elicitation stage (the evidence experienced over time; 4 levels), scenario (color / high base rate, and number / low base rate; 2 levels), observed vs predicted (participant data vs BIBN model prediction; 2 levels), and the between-subject condition of scenario order (color first, number first; 2 levels). This analysis revealed that scenario order had no effect on estimates, $BF_{Inclusion} = 0.023$, whilst estimates increased significantly across elicitation stages (T3 > T2 > T1 > Baseline, Fig. 4), $BF_{Inclusion} > 10000$, were higher in the color (vs number) scenario (grey vs. black lines, Fig. 4), $BF_{Inclusion} > 10000$, and were higher in BIBN model predictions than in participant estimates (dashed vs. solid lines, Fig. 4), $BF_{Inclusion} > 10000$. The significant interaction of scenario and elicitation stage, $BF_{Inclusion} >$

10000, revealed estimates to increase significantly more in the number scenario. There were also significant interactions of scenario and observed vs predicted, $BF_{Inclusion} > 10000$, elicitation stage and observed vs predicted, $BF_{Inclusion} > 10000$, and crucially, scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion} > 10000$. These reveal participants to be a) generally insufficient in their updating over time, and b) that this insufficiency is worse in the number (vs color) scenario. The model that included the above significant terms yielded the most significant model improvement (all other terms were not significant), $BF_M = 932.54$, and was decisive overall, $BF_{10} > 10000$.

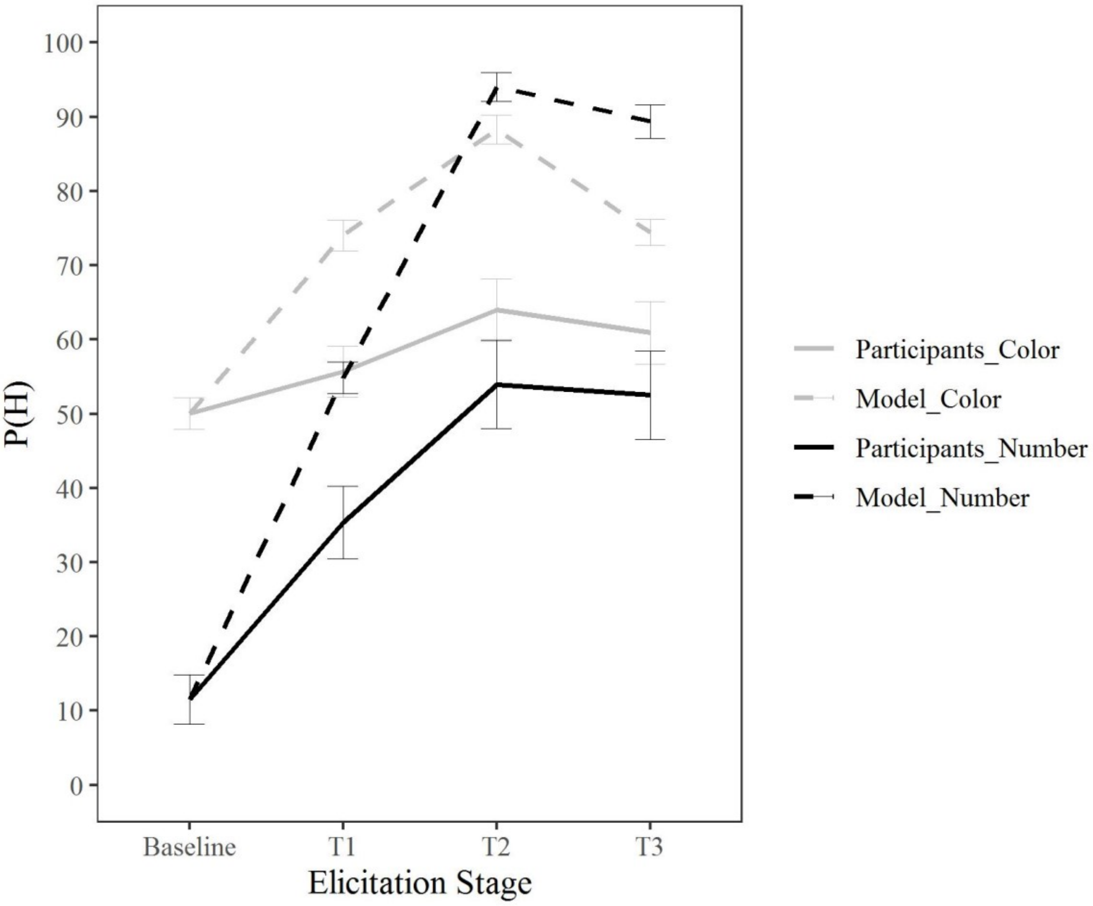

Fig. 4.

Probability of hypothesis estimates across evidence stages, black lines represent number scenario, grey lines represent color scenario. Dashed lines are BIBN model predictions (fitted on individual level), whilst solid lines are participant responses. Error bars reflect 95% Confidence Intervals.

To determine whether this insensitivity to evidence a) persisted across both trends of corroborating and contradicting reports, and b) whether this was again exacerbated by base rate differences (effect of scenario), two further Bayesian repeated measures ANOVA were conducted on a split of elicitation stages; Baseline to T2 (corroborating trend), and T2 to T3 (contradicting trend).

As in the overall analysis, we find the same significant trends in the Baseline to T2 corroborating trend. More precisely, again color estimates are generally higher than number (grey vs. black lines, Fig. 4), $BF_{Inclusion}$ > 10000, predicted higher than observed (dashed vs. solid line, Fig. 2), $BF_{Inclusion}$ > 10000, and estimates increase across elicitation stages (T2 > T1 > Baseline, Fig. 2), $BF_{Inclusion}$ > 10000. Critically, we once again find significant interactions of scenario and elicitation stage, $BF_{Inclusion}$ > 10000, scenario and observed vs predicted, $BF_{Inclusion}$ = 3187, elicitation stage and observed vs predicted, $BF_{Inclusion}$ > 10000, and the interaction of scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion}$ = 1549. Taken together, these show that deviations from predicted are greater in number (vs color) scenarios, greater across elicitation stages, and that this increasing deviation is greater in number (vs color) scenarios. Once again, the model including these terms yielded the most significant model improvement, BF_M = 1548.69, and was decisive overall, BF_{10} > 10000.

Similarly, in the T2 to T3 (contradicting) trend, we find significant effects of scenario (color > number, grey vs. black lines, Fig. 4), $BF_{Inclusion}$ > 10000, elicitation stage (T3 < T2), $BF_{Inclusion}$ > 10000, and observed vs predicted (predicted > observed, dashed vs. solid lines, Fig. 4), $BF_{Inclusion}$ > 10000. Crucially, we also find the same significant interaction terms of scenario and elicitation stage, $BF_{Inclusion}$ = 4.59, scenario and observed vs predicted, $BF_{Inclusion}$ > 10000, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 26.62, and the interaction of scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion}$ = 4.85. Taking these together, we demonstrate that participants are generally insufficient in their downward adjustment of probability given a contradicting report, but that the insufficiency of this decrease is greater in the color (vs number) scenario. Finally, the model including the above terms yielded the most significant model improvement, $BF_M = 13.12$, and was decisive overall, $BF_{10} > 10000$.

Overall, we find support for all three hypotheses, wherein we not only find insufficient updating for both the introduction of confirmatory and contradictory reports, but also that this insufficiency is exacerbated in the low base rate (number) scenario, such that the impact of the conjunction of low base rates and corroboration is severely *underestimated*.

3.2.2. Probability of reliability of first witness

Participants were asked to estimate the probability of the first witness being reliable across 4 time-points, at baseline (no reports made), T1 (only the first witness has reported), T2 (the first witness report has now been corroborated by the second witness), and T3 (the third witness has contradicted the first and second witness). Consequently, it is of interest whether participants appropriately update the reliability of this witness over time. More precisely, whether participants 1) incorrectly update the reliability of the first witness when only that witness has reported (T1), 2) increase the reliability of the first witness sufficiently given the corroboration provided by the second witness (T2), and 3) decrease the reliability of the first witness appropriately given the contradicting report of the third witness (T3). As in the P(Hypothesis) dependent variable, it is of critical interest whether participants are sensitive to the impact of the base rate (number / low base rate, color / high base rate) on reliability, wherein reliability should be updated to a greater degree given the coherence between corroborating rare reports, which should in turn be resistant to a contradicting (minority) report.

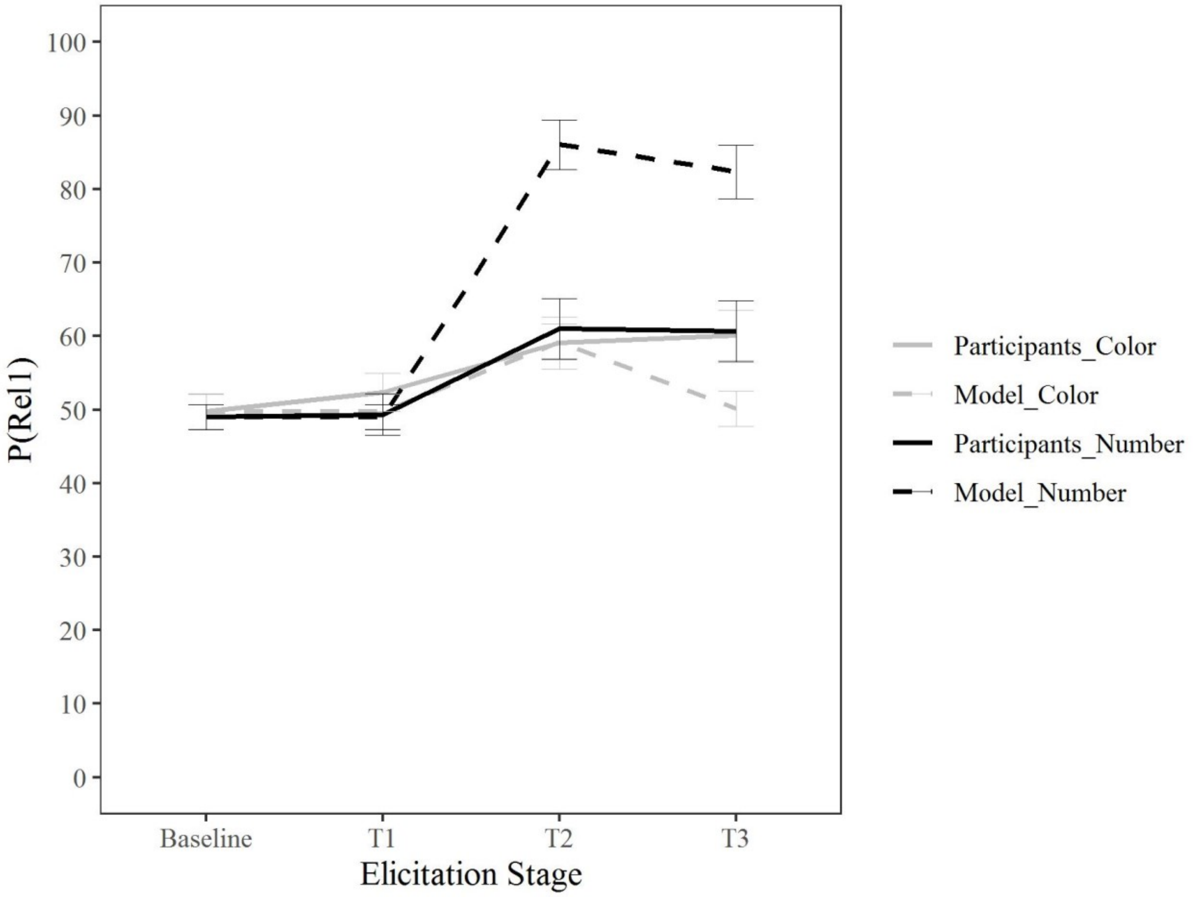

Fig. 5.

P(Rel$_1$) estimates across evidence stages, black lines represent number scenario, grey lines represent color scenario. Dashed lines are BIBN model predictions (fitted on individual level), whilst solid lines are participant responses. Error bars reflect 95% Confidence Intervals.

To test hypotheses 1, 2, and 3, analyses need to be split by elicitation stage pairings (1: Baseline to T1, 2: T1 to T2, and 3: T2 to T3). However, first it is worth noting that the overall Bayesian repeated measures ANOVA including all factors corresponds to the findings of P(Hypothesis) data. More precisely, strong support for the null was found for order of scenarios $BF_{Inclusion}$ = 0.042, whilst significant effects were found for scenario (number > color, black vs. grey lines, Fig. 5), $BF_{Inclusion}$ > 10000, elicitation stage (increasing overall trend, T3 > T2 > T1 > Baseline, Fig. 5), $BF_{Inclusion}$ > 10000, and observed vs predicted (predicted > observed, dashed vs. solid lines, Fig. 5), $BF_{Inclusion}$ > 10000. Again, we also find the same significant interaction terms of scenario and elicitation stage, $BF_{Inclusion}$ > 10000, scenario and observed vs predicted, $BF_{Inclusion}$

> 10000, elicitation stage and observed vs predicted, $BF_{Inclusion}$ > 10000, and the interaction of scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion}$ > 10000. Taken together, this again demonstrates participant insensitivity to evidence across updates, and effect that is exacerbated in the low base rate (number) scenario. Consequently, the model including the above significant terms was both the best fit, $BF_M = 503.13$, and decisive overall, $BF_{10} > 10000$.

Baseline to T1. To test hypothesis 1 - whether participants are erroneously double-counting the impact of a single report (by updating reliability as well as the likelihood of the hypothesis) - a Bayesian repeated measures ANOVA was run on the baseline to T1 elicitation stages only. In accordance with the flat lines of Baseline to T1 in Fig. 5, there was strong evidence for the null across all factors (scenario, $BF_{Inclusion} = 0.32$; elicitation stage, $BF_{Inclusion} = 0.062$; observed vs predicted, $BF_{Inclusion} = 0.062$) and interaction terms. This demonstrates that participants are correctly keeping reliability estimates constant, given a single reporter.

T1 to T2. To test the second hypothesis – whether the introduction of a corroborating report leads to insufficient reliability updating (and whether this is exacerbated when base rates are low) – a Bayesian repeated measures ANOVA was run on the T1 to T2 subset of elicitation stages. Along with main effects of scenario (number > color, black vs. grey lines, Fig. 5), $BF_{Inclusion}$ > 10000, elicitation stage (T2 > T1), $BF_{Inclusion}$ > 10000, and observed vs predicted (predicted > observed, dashed vs. solid lines, Fig. 5), $BF_{Inclusion}$ > 10000, significant interactions were found for scenario and elicitation stage, $BF_{Inclusion}$ > 10000, scenario and observed vs predicted, $BF_{Inclusion}$ > 10000, elicitation stage and observed vs predicted, $BF_{Inclusion}$ > 10000, and scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion}$ > 10000. The model with the above terms yielded the most significant fit, $BF_M > 10000$, and was decisive overall, BF_{10} > 10000. Taking these results together, there is evidence for generally insufficient updating given the corroborating report, but this effect is driven by the larger deviation in the low base rate (number) scenario. To corroborate this difference, separate Bayesian repeated measures ANOVA were run on the color and number scenarios in isolation. These found no deviation from normative expectation across elicitation stages in the color scenario (T1 to T2, grey solid vs grey dashed lines, Fig. 5), $BF_{Inclusion} = 0.24$, but decisive evidence for such a deviation across elicitation stages in the number scenario, $BF_{Inclusion} > 10000$, in line with the above interaction.

T2 to T3. To test the third hypothesis – whether the introduction of a contradicting (minority) report leads to an appropriate reduction in reliability estimates – a final Bayesian repeated measures ANOVA was run on the T2 to T3 subset of elicitation stages. Here we again find main effects of scenario (number > color, black vs. grey lines, Fig. 5), $BF_{Inclusion} > 10000$, elicitation stage (T3 < T2), $BF_{Inclusion} = 137.36$, and observed vs predicted (predicted > observed, dashed vs. solid lines, Fig. 5), $BF_{Inclusion} > 10000$. There were significant interactions between scenario and observed vs predicted, $BF_{Inclusion} > 10000$, and elicitation stage and observed vs predicted, $BF_{Inclusion} = 77.25$, which indicate that participants are a) further from normative expectation in the low base rate (number) scenario, and b) do not adjust appropriately to the new, contradicting evidence. However, no evidence was found for the interaction of scenario and elicitation stage, $BF_{Inclusion} = 0.56$, and scenario, elicitation stage and observed vs predicted, $BF_{Inclusion} = 1.53$, indicating that the overall decrease in estimates – whether split by observed vs prediction, or not – across elicitation stages did not significantly differ across scenarios. Lastly, the model only including the above significant terms providing the most significant fit, $BF_M = 60.19$, and decisive overall, $BF_{10} > 10000$. To understand these interactions, it is worth noting that when splitting the analysis by scenario, in the color scenario there is a significant evidence x observed vs predicted interaction, $BF_{Inclusion} = 3817$, whilst there was no such interaction in the number scenario, $BF_{Inclusion} = 0.21$. This shows that when base rates are high (color scenario), participants should penalise reliability more given a contradicting report, whilst when base rates are low (number scenario) the corroborating majority should retain their reliability (despite the contradicting minority report), which tracks the (albeit generally insufficient estimates of reliability) trend in participants.

Taken together, we show that participants correctly understand the inability to update reliability without corroboration (hypothesis 1), but fail to update sufficiently given a corroborating report – when low base rates dictate a more substantial increase (hypothesis 2). Finally, we show that participants fail to appreciate the appropriate penalization of reliability when a contradicting (high base rate) report is introduced (hypothesis 3).

3.2.3. Probability of reliability of third witness

Having analysed the reliability of witness 1 across elicitation stages, the final comparison of which entails witness 1 having been corroborated by one witness, and contradicted by a third, we next turn to estimates of the reliability of this final, contradicting witness – drawing it in comparison to the corroborating majority. To assess this, a Bayesian repeated measures ANOVA was conducted in the same manner as previous dependent variables, with the sole exception of replacing the elicitation stage factor with a witness (1 vs 3; 2 levels) within-subject factor.

Table 4.

Reliability estimate comparison of witness 3 (contradicting minority) and witness 1 (corroborated majority) across scenarios.

		Observed			Predicted		
Scenario	Witness	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
Color	1 (corroborated majority)	60.1	18.85	118	50.12	13.19	118
	3 (contradicting minority)	41.02	19.32	118	15.59	3.43	118
Number	1 (corroborated majority)	60.69	22.72	118	82.35	19.88	118
	3 (contradicting minority)	41.89	24.83	118	5.14	4.93	118

As previously, this analysis found no evidence of a scenario order effect, $BF_{Inclusion} = 0.034$, but did find a significant effect of witness (witness 3 < witness 1; confirming the manipulation worked), $BF_{Inclusion} > 10000$, scenario (color < number), $BF_{Inclusion} > 10000$, and observed vs predicted (observed > predicted), $BF_{Inclusion} > 10000$. The significant interactions of scenario and witness, $BF_{Inclusion} > 10000$, scenario and observed vs predicted, $BF_{Inclusion} > 10000$, witness and observed vs predicted, $BF_{Inclusion} > 10000$, and the three-way interaction of scenario, witness, and observed vs predicted, $BF_{Inclusion}$ > 10000, reveal participant insensitivity to a number of factors.[5] More precisely, participants significantly overestimate the reliability of the contradicting witness, an effect that is exacerbated by the severe penalty to reliability in the low base rate (number) scenario. Conversely, participants underestimate the high reliability of corroborating low base rate (number) witnesses, whilst overestimating the penalised reliability of corroborating high base rate (color) witnesses. Together, these reflect a failure to sufficiently update reliability in both contradicting and corroborating witnesses, and further, a failure to account for the impact of base rates – leading to both exacerbated under (corroborating) and over (contradicting) estimation.

3.2.4. Probability of third witness report

The final dependent variable of interest are the predictive estimates of how likely the third witness is to make a confirmatory (ball is red / ball is 34) report. These estimates were elicited at two time-points: when only witness 1 has made a confirmatory report, and when witness 2 has corroborated that report. There are two questions of interest. First, whether participants understand qualitatively that prediction likelihood should increase as more corroborative reports come in (i.e. a manipulation check). Secondly, as in the preceding dependent variables, whether participants understand the implications of base rates and corroboration, but in this case on predictive reasoning. To address these questions, a Bayesian repeated measures ANOVA was conducted, incorporating all the factors of preceding analyses (albeit with elicitation stage restricted to time-points T1 and T2).

[5] The model including all the above significant terms yielded the most significant fit, $BF_M = 627.31$, and was decisive overall, $BF_{10} > 10000$.

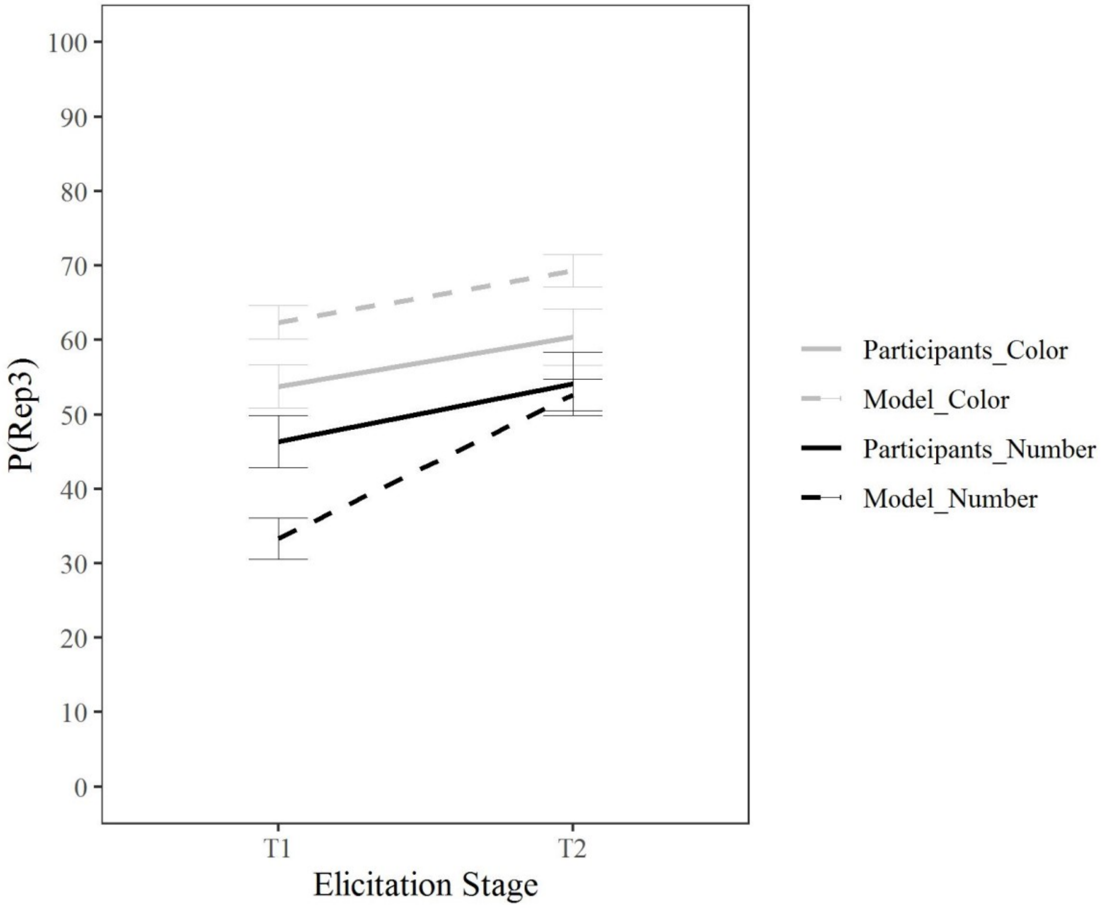

Fig. 6.

P(Rep3) estimates across evidence stages, black lines represent number scenario, grey lines represent color scenario. Dashed lines are BIBN model predictions (fitted on individual level), whilst solid lines are participant responses. Error bars reflect 95% Confidence Intervals.

As in previous analyses, strong evidence for a null effect of scenario order was found, $BF_{Inclusion}$ = 0.093, whilst significant effects were found for scenario (color > number, grey vs. black lines, Fig. 6), $BF_{Inclusion}$ > 10000, observed vs predicted (predicted > observed, dashed vs. solid lines, Fig. 6), $BF_{Inclusion}$ > 10000, and critically, elicitation stage (T2 > T1), $BF_{Inclusion}$ > 10000. The significant interactions of scenario and elicitation stage, $BF_{Inclusion}$ = 1123.16, scenario and observed vs predicted, $BF_{Inclusion}$ > 10000, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 222.15, and scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion}$ = 85.45, speak to the second hypothesis.[6] More precisely, whilst the high base rate (color) scenario participant prediction estimates approximately track the change expected by BIBN model predictions, participants are insensitive to the lower starting point – and more substantial change in predictive likelihood – entailed by the low base rate (number) scenario. This again reflects an insensitivity among participants to the impact of base rates in conjunction with corroborative testimony.

3.3. Conclusion

Replicating the findings of Experiment 1, we find once again that although lay reasoners appreciate the general influence of corroboration (extending this to understanding of reliability updates among corroborating witnesses, and prediction of subsequent reports), they remain insensitive to the critical role of hypothesis base rates. More precisely, extending previous findings to probability estimates (and using a novel scenario based on roulette), we find substantial underestimation in diagnostic inferences when base rates are rare.

4. Experiment 3

Experiment 3 is designed to replicate Experiment 2, with the exception of using non-human witnesses. This was done to encourage participants to treat witnesses as independent of each other. As in Experiment 2, participants are presented with two scenarios of a semi-blind roulette game: in the color scenario, they want to know whether the ball has landed on red or black (P(H)= .5); in the number scenario, they want to know whether the ball has landed on a specific number (P(H) ≈ .03). Participants cannot observe the spin outcomes themselves and have to rely on reports that come from three robots (the non-human witnesses). The robots provide reports independently; each robot is reliable 50% of the time; when unreliable it will report the outcome randomly. Each robot provides one report, and participants are asked to make judgments

[6] The model including the above significant terms yielded the strongest fit, BF_M = 189.07, and was decisive overall, BF_{10} > 10000.

following each one. Therefore, the Bayesian Networks that illustrate these two scenarios are the same as those used in Experiment 2.

The hypotheses remain the same as in Experiment 2:

1. Reasoners may erroneously "double count" a single reporting (non-human) witness, such that not only is the probability of the hypothesis updated, but the probability of the witness being reliable is also (falsely) updated. We note however, that this hypothesis had a significant null effect in Experiment 2, but here we seek to confirm this via replication.
2. There will be an overall underestimation in belief updating, across all sequential evidence presentations (irrespective of hypothesis base rates).
3. Lastly, we predict that reasoners will substantially *underestimate* the posterior probabilities, given corroborative reports, of a) the reported hypothesis, b) the reliability of corroborating witnesses, and c) the likelihood of further corroborating reports when *base rates are lower* (number scenario) rather than higher (color scenario) – i.e. an inversion of standard base rate neglect.

4.1. Method

Participants. 120 participants were recruited and participated online through MTurk (https://www.mturk.com/). Those eligible for participation had a 95% and above approval rating from over 100 prior HITs, and could not have taken part in previously studies. Participants were English speakers, located in the United States. From the 120 participants, 2 were removed for incomplete data, and 4 for not meeting the language and location requirements. Of the 114 remaining participants, 50 were female. The mean age was 35.73 (*SD* = 11.42). Participants gave informed consent and were paid $1 for their time (*Median* = 6.73 minutes, *SD* = 2.68).

Materials and Procedure. As in the second experiment, participants are presented with two roulette scenarios: a high base rate scenario is the color scenario (the ball has landed on red) and the low base rate scenario is the number scenario (the ball has landed on no. 34). Participants cannot observe the spin outcomes themselves and have to rely on reports provided by three robots. The robots provide reports independently of each other. Each robot uses a camera to record spin outcomes, and the signal from the camera is transformed by a processor that lights up a bulb to indicate the report (red/black in the color scenario; 1-36 in the number scenario). The signal from the camera is sometimes scrambled – it is only reliable 50% of the time. When the signal is unreliable, the processor lights up one of the bulbs *randomly*. When the signal is reliable, the processor always lights up the correct bulb. The order of scenarios was counterbalanced between participants. Each robot provides a single report, and participants are view these one at a time. The first and the second report support the hypothesis (The ball has landed on red/ 34), and the third report contradicts the hypothesis (the ball has not landed on red/ The ball has landed on another number (not 34)). Participants provide estimates of the hypothesis before seeing any reports (baseline), after the first report (T1), after the second report (T2) and after the third report (T3).

At Baseline, prior to seeing any reports, participants were asked to provide two estimates: 1) the probability that the ball has landed on red/34, 2) the probability of each witness/robot being reliable. Subsequently, participants were asked to update their beliefs based on the evidence received. After each of the three reports, the following two estimates were required:

Probability of the Hypothesis: *What is the probability that the ball has landed on red/34, given everything you know so far?*

Reliability of the First Robot: *What is the probability that robot 1/A is reliable (i.e. paid attention to this outcome); given everything you know so far?*

After the first two reports, there was also a third question that required participants to make a Third Robot Prediction:

What is the probability that robot 3/C will also report red/34?

After the third report, the third question required participants to make an estimate of Reliability of the Third Robot:

What is the probability that robot 3/C is reliable (i.e. paid attention to this outcome); given everything you know so far?

At the end of each question, participants were reminded that the estimate being asked for refers to the particular instance described (based on the evidence received), and not the general likelihood (e.g., the probability of the ball landing on red in general). For a full description of the methods, see Supplementary Materials.

4.2. Results

As in previous experiments, all analyses were Bayesian, and elicited prior probabilities for the hypothesis (ball landed on red/34) as well as witness reliability from each participant, were used to fit individual Bayesian Networks (BIBNs). As previously, these fitted models accommodate for a) participants bringing their own adjustments to the priors stated in the materials, and b) the influence of said assumptions on subsequent model predictions. The remaining structure and parameters were taken from the background information presented to all participants. The models were then used to generate normative predictions (probability estimates) for each participant across elicitation stages, to be used in subsequent comparison analyses.

4.2.1. Probability of the hypothesis

Our hypotheses regarding participant estimates of the hypotheses (ball has landed on red/ ball has landed on no. 34) remain the same. The first hypothesis predicts that participants will be generally conservative in their belief updating, throughout the experiment. Our second hypothesis is that conservatism in belief updating will occur both when reports corroborate and contradict each other. Our third hypothesis is that participants will underestimate the impact of evidence to a larger extent when the base rate is low rather than high (they will fail to appreciate the impact of corroboration in the low base rate scenario).

A Bayesian repeated measures ANOVA was used to investigate whether participants are generally conservative in their belief updating. The analysis included the following factors: elicitation stage (baseline, T1, T2 and T3; 4 levels); scenario (color scenario/ number scenario; 2 levels), observed vs predicted (participant data vs BIBN model prediction; 2 levels), and the between-participants factor of scenario order (color first, number first; 2 levels). The scenario order did not have a significant effect on estimates, $BF_{Inclusion} = 0.195$. There was a significant effect of elicitation stage as estimates increased across the elicitation stages (Fig. 7), $BF_{Inclusion} > 10000$, estimates were higher in the color scenario than in the number scenario (grey vs. black lines, Fig. 7),), $BF_{Inclusion} > 10000$, and BIBN model predictions were higher than participant estimates, (dashed vs. solid lines, Fig. 7), $BF_{Inclusion} > 10000$. Interactions between the following factors were also found to be significant: scenario and observed vs predicted, $BF_{Inclusion} > 10000$, elicitation stage and observed vs predicted, $BF_{Inclusion} > 10000$, and scenario, elicitation stage and observed vs predicted, $BF_{Inclusion} > 10000$. These interactions reflect conservatism in participant estimates throughout the experiment, which diverge from the BIBN model predictions, more so in the number than the color scenario. The model that included the above significant terms yielded the most significant model improvement, $BF_M = 109.08$, and was decisive overall, $BF_{10} > 10000$.

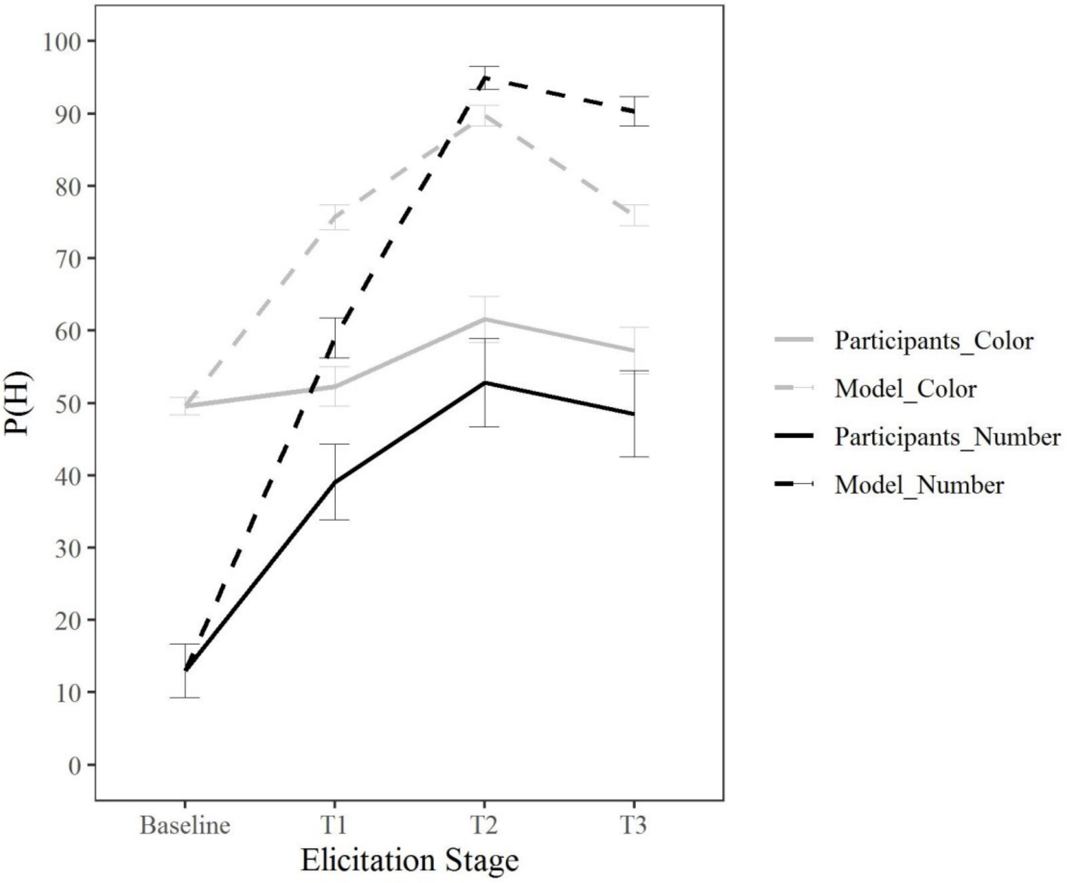

Fig. 7.

Probability of hypothesis estimates across evidence stages, black lines represent number scenario, grey lines represent color scenario. Dashed lines are BIBN model predictions (fitted on individual level), whilst solid lines are participant responses. Error bars reflect 95% Confidence Intervals.

In order to investigate whether participants are conservative with both corroborating and contradicting evidence, and whether this effect was higher in the number than in the color scenario, two Bayesian repeated measures ANOVA were done, split by elicitation stages as follows: Baseline to T2 (corroborating reports), and T2 to T3 (contradicting reports).

The analysis of estimates from Baseline to T2 shows that estimates are higher in the color (grey vs. black lines, Fig. 7), $BF_{Inclusion}$ > 10000, BIBN model predictions are higher than participant estimates, (dashed vs. solid line, Fig. 7), $BF_{Inclusion}$ > 10000, and follow an upward trend across elicitation stages (Fig. 7), $BF_{Inclusion}$ > 10000. Critically, there were significant interactions between the following factors: scenario and elicitation stage, $BF_{Inclusion}$ > 10000, scenario and observed vs predicted, $BF_{Inclusion}$ = 216.05, elicitation stage and observed vs predicted, $BF_{Inclusion} > 10000$, and scenario, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 1145.56. These interactions reveal that, as estimates increase across stages, participant estimates diverge from BIBN model predictions more so in the number than in the color scenario. The model including these terms yielded the most significant model improvement, BF_M = 1145.56, and was decisive overall, $BF_{10} > 10000$.

The analysis of estimates from T2 to T3 (the contradicting reports) revealed that estimates were higher in the color scenario (grey vs. black lines, Fig. 7), $BF_{Inclusion} > 10000$, estimates were lower at T3 than T2, $BF_{Inclusion} > 10000$, and BIBN model predictions were higher than participant estimates, (dashed vs. solid lines, Fig. 7), $BF_{Inclusion} > 10000$. The interaction between elicitation stage and scenario was not significant, $BF_{Inclusion}$ = 2.241. Significant interactions were found between scenario and observed vs predicted, $BF_{Inclusion} > 10000$, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 3.33, and scenario, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 3.70. The analysis shows that participants fail to show the downward trend that is more pronounced in the color than in the number scenario. The model including the above significant terms yielded the most significant model improvement, BF_M = 6.33, and was decisive overall, $BF_{10} > 10000$.

The results of analyses show support for our hypotheses: 1) there is insufficient adjustment with both corroborating and contradicting reports, and 2) this conservatism is more pronounced in the number than in the color scenario. As such, once again participants fail to appreciate the effect of corroboration, and its increased impact in instances of low base rates.

4.2.2. Probability of reliability of first witness

Estimates of reliability of the first witness were elicited at all four stages. At baseline, no reports are made; at T1 there is a report from the first witness; and T2 there is a corroborating report from the second witness; at T3 there is a contradicting report from the third witness. Corroboration has a larger effect on reliability of the first witness in the number than in the color scenario. Following corroboration, a contradicting report has a much smaller effect in the number than in the color scenario. We predict that participants will fail to appreciate the impact of base rate in the number scenario. The analysis of estimates of reliability of the first witness will investigate a) whether participants incorrectly update their estimates at T1, b) are conservative in their belief updating after corroboration occurs (at T2), and c) decrease their estimates normatively following a contradicting report (at T3).

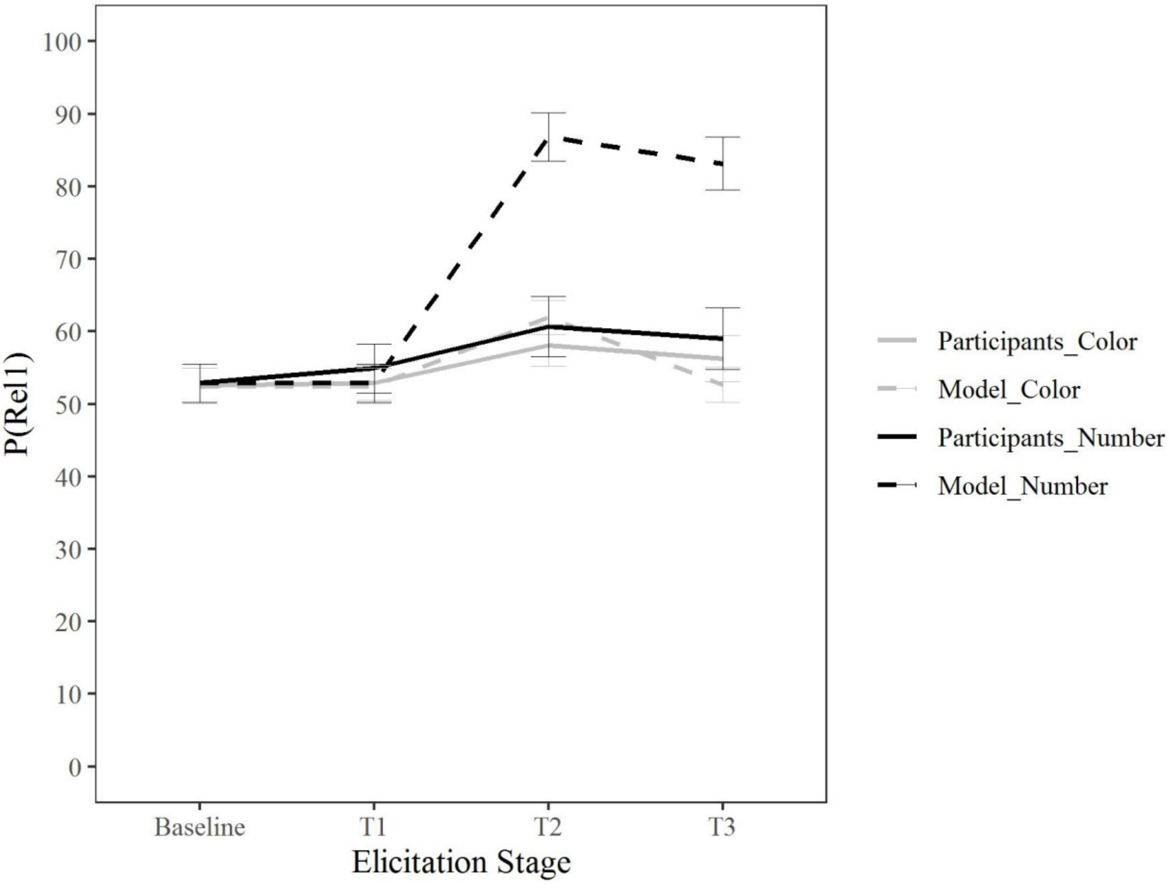

Fig. 8.

P(Rel_1) estimates across evidence stages, black lines represent number scenario, grey lines represent color scenario. Dashed lines are BIBN model predictions (fitted on individual level), whilst solid lines are participant responses. Error bars reflect 95% Confidence Intervals.

The overall Bayesian repeated measures ANOVA revealed that the main effect of scenario order was not significant, $\text{BF}_{\text{Inclusion}} = 0.147$. Estimates were higher in the number scenario, (black vs. grey lines, Fig. 8), $\text{BF}_{\text{Inclusion}} > 10000$, increased across elicitation stages (Fig. 8), $\text{BF}_{\text{Inclusion}} >$

10000, and BIBN model predictions were higher than participant estimates (dashed vs. solid lines, Fig. 8), $BF_{Inclusion} > 10000$. There are significant interactions between scenario and elicitation stage, $BF_{Inclusion} > 10000$, scenario and observed vs predicted, $BF_{Inclusion} > 10000$, elicitation stage and observed vs predicted, $BF_{Inclusion} > 10000$, and the interaction of scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion} > 10000$. Participants underestimate the impact of evidence across elicitation stages, and this deviation from normative predictions is larger in the number scenario. The model including the above significant terms was both the best fit, $BF_M = 144.64$, and decisive overall, $BF_{10} > 10000$.

In order to investigate whether conservatism occurs after corroboration, as well as a contradicting report, separate analyses are done on estimates at baseline to T1, T1 to T2, and T2 to T3.

Baseline to T1. A Bayesian repeated measures ANOVA was run on estimates at baseline and T1 to investigate whether participants incorrectly update reliability after the first report. The effects of scenario, $BF_{Inclusion} = 0.061$, elicitation stage, $BF_{Inclusion} = 0.041$, and observed vs predicted, $BF_{Inclusion} = 0.045$, were not significant. Interactions between these factors were also not significant. Therefore, we once again find that participants do not update reliability after the first report.

T1 to T2. Another Bayesian repeated measures ANOVA was run on T1 to T2 estimates to investigate whether participants underestimate the impact of corroboration, which is larger in the number scenario. Estimates were higher in the number scenario (black vs. grey lines, Fig. 8), $BF_{Inclusion} > 10000$, higher at T2 than T1, $BF_{Inclusion} > 10000$, and higher in BIBN model predictions (dashed vs. solid lines, Fig. 8), $BF_{Inclusion} > 10000$. There were significant interactions between scenario and elicitation stage, $BF_{Inclusion} > 10000$, scenario and observed vs predicted, $BF_{Inclusion} > 10000$, elicitation stage and observed vs predicted, $BF_{Inclusion} > 10000$, and scenario, elicitation stage, and observed vs predicted, $BF_{Inclusion} > 10000$. The model with the above terms yielded the most significant fit, $BF_M > 10000$, and was decisive overall, $BF_{10} > 10000$. While participants underestimate the impact of evidence, this effect appears to be driven by insufficient updating in the number scenario. Separate analyses on estimates in each scenario confirm this suggestion. Participant estimates were not different from BIBN model predictions in the color scenario (T1 to T2, grey solid vs grey dashed lines, Fig. 8), $BF_{Inclusion}$ = 2.16, whereas this difference was observed in the number scenario, $BF_{Inclusion}$ > 10000.

T2 to T3. A Bayesian repeated measures ANOVA was run on T2 to T3 estimates to test whether participants follow an appropriate downward trend after a contradicting report. Estimates were higher in the number scenario (black vs. grey lines, Fig. 8), $BF_{Inclusion}$ > 10000, were higher at T2 than T3, $BF_{Inclusion}$ = 495.75, and BIBN model predictions were higher than participant estimates (dashed vs. solid lines, Fig. 8), $BF_{Inclusion}$ > 10000. There were significant interactions between scenario and observed vs predicted, $BF_{Inclusion}$ > 10000, and elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 4.29, but the interactions between scenario and elicitation stage, $BF_{Inclusion}$ = 0.82, and scenario, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 0.92, were not significant. These results show that estimates in the two scenarios were not significantly different overall, but there was significant deviation from normative estimates after a contradictory report, and this deviation was larger in the number than in the color scenario. The model only including the above significant terms providing the most significant fit, BF_M = 15.63, and decisive overall, BF_{10} > 10000.

A significant interaction between elicitation stage and observed vs predicted was found in the color scenario, $BF_{Inclusion}$ = 16.96, but not the number scenario, $BF_{Inclusion}$ = 0.292. A significant interaction in the color scenario reflects a downward trend in BIBN model estimates, which is not observed in participant estimates. Conversely, the deviation between the BIBN model estimates and participant estimates remained constant across elicitation stages in the number scenario.

The results show that participant estimates of reliability remain appropriately constant after the first report (hypothesis 1); are insufficiently high following corroboration in the number scenario (hypothesis 2), and do not follow normative downward trends after a contradicting report in the color scenario (hypothesis 3).

4.2.3. Probability of reliability of third witness

Estimates of reliability of the third witness are taken after the two corroborating reports are contradicted by the third (minority) report (at T3), and are investigated in comparison to estimates of reliability of the first witness (also at T3). A Bayesian repeated measures ANOVA was conducted where a witness (1 vs 3, 2 levels, within-subjects) replaces elicitation stage as a factor (all of the other previously described factors remain the same).

Table 5.

Reliability estimate comparison of witness 3 (contradicting minority) and witness 1 (corroborated majority) across scenarios.

		Observed			Predicted		
Scenario	Witness	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
Color	1 (corroborated majority)	56.25	17.12	114	52.63	12.87	114
	3 (contradicting minority)	46.14	17.96	114	15.40	3.33	114
Number	1 (corroborated majority)	59.03	22.84	114	83.15	19.71	114
	3 (contradicting minority)	45.46	23.99	114	5.14	5.27	114

The main effect of scenario order was not significant, $BF_{Inclusion} = 0.021$. Estimates of reliability of witness 3 were (appropriately) lower than those of witness 1, $BF_{Inclusion} > 10000$, were higher in the number scenario, $BF_{Inclusion} > 10000$, and participant estimates were higher than BIBN model predictions, $BF_{Inclusion} > 10000$. There were significant interactions between scenario and witness, $BF_{Inclusion} > 10000$, scenario and observed vs predicted, $BF_{Inclusion} > 10000$, witness and observed vs predicted, $BF_{Inclusion} > 10000$, and scenario, witness and observed vs predicted, $BF_{Inclusion} > 10000$. The model including all the above significant terms yielded the most significant fit, $BF_M = 991.90$, and was decisive overall, $BF_{10} > 10000$. The results show that participants overestimate the reliability of a contradicting minority (the third witness) in both scenarios, although this deviation is larger in the number scenario. Participants overestimate reliability of the first (corroborated witness) in the color scenario, but underestimate it in the number scenarios. This finding confirms that participant fail to appreciate the impact of corroboration (and conversely, contradiction) on reliability of witnesses generally, as well as the sensitivity of this effect to base rates.

4.2.4. Probability of third witness report

Probability of third witness report refers to predictive estimates of the likelihood that the third report supports the hypothesis (ball has landed on red/ball has landed on 34). Participants provide these estimates after the first and the second (corroborating) report. These estimates were analysed to investigate whether participants follow a normative (upward) trend after the second report, and whether they understand that the increase is larger in the low base rate scenario. Estimates were analysed using a Bayesian repeated measures ANOVA that included all of the factor previously described (elicitation stage has only 2 levels: T1 and T2).

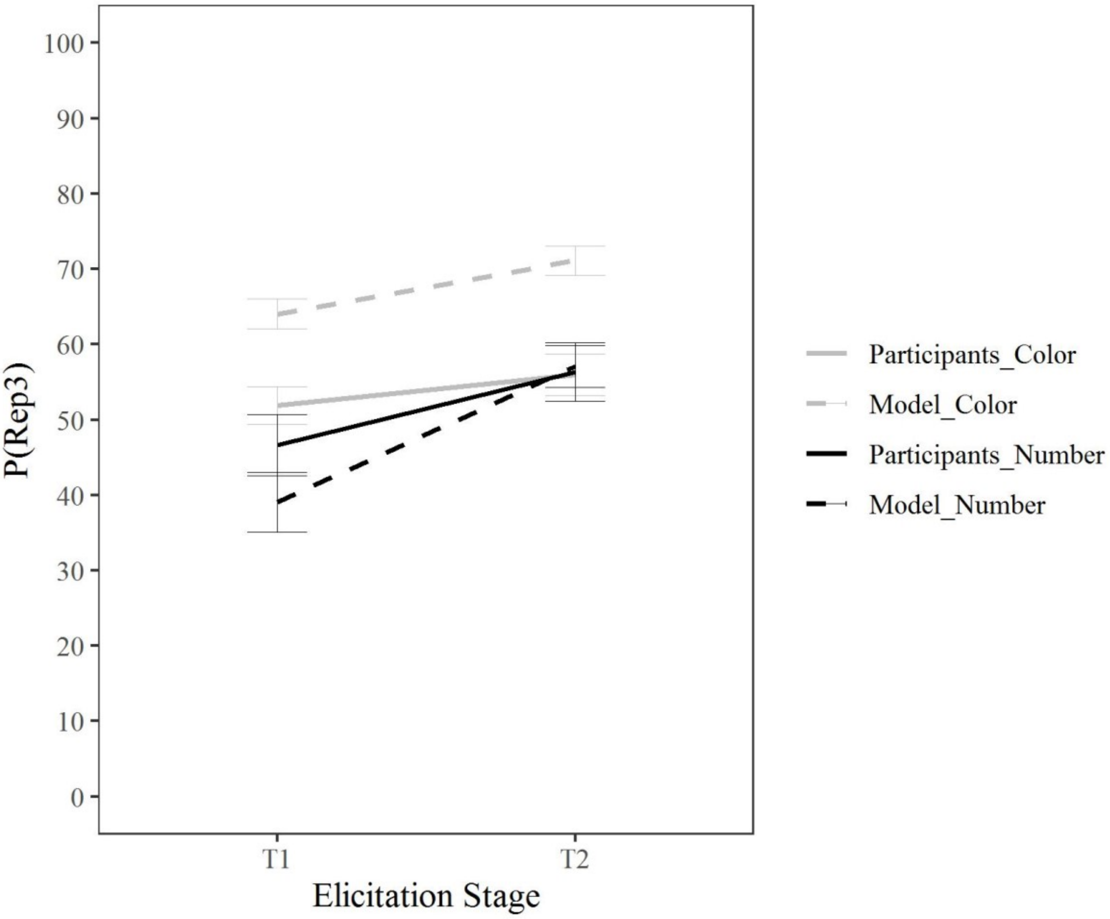

Fig. 9.

P(Rep_3) estimates across evidence stages, black lines represent number scenario, grey lines represent color scenario. Dashed lines are BIBN model predictions (fitted on individual level), whilst solid lines are participant responses. Error bars reflect 95% Confidence Intervals.

The main effect of scenario order was not significant, $BF_{Inclusion}$ = 0.092. Estimates were higher in the color scenario (grey vs. black lines, Fig. 9), $BF_{Inclusion}$ > 10000, BIBN model predictions were higher than participant estimates (dashed vs. solid lines, Fig. 9), $BF_{Inclusion}$ > 10000, and higher at T2 than T1, $BF_{Inclusion}$ > 10000. There were significant interactions between scenario and elicitation stage, $BF_{Inclusion}$ = 1440.61, scenario and observed vs predicted, $BF_{Inclusion}$ > 10000, elicitation stage and observed vs predicted, $BF_{Inclusion}$ = 10.75. The three way interaction between scenario, elicitation stage, and observed vs predicted was not significant, $BF_{Inclusion}$ = 2.26. The model including the above significant terms yielded the strongest fit, BF_M = 111.61, and was decisive overall, BF_{10} > 10000. Participants do not follow normative predictions across both scenarios, although the increase in the color scenario is approximately parallel to BIBN model predictions. Participants do not appreciate the low base rate in the number scenario, and the increase that follows after corroboration.

4.3. Conclusion

Replicating the findings of both Experiment 1 and Experiment 2, we find once again that although lay reasoners appreciate the general influence of corroboration (extending this to the example of non-human witnesses), they remain insensitive to the influence of hypothesis base rates.

5. General Discussion

Across two experiments we employed a novel paradigm in which we directly manipulate the base rate of a reported hypothesis - whether in a more qualitative burglary case (Experiment 1), or within a constrained, more quantitative roulette scenario involving human (Experiment 2) and non-human (Experiment 3) witnesses - looking at the influence of corroboration (and contradiction) within a sequence of diagnostic reasoning. Whilst Experiment 1 illustrated the central error under investigation in qualitative terms, Experiments 2 and 3 then both replicated and enriched this finding, by placing it in the context of surrounding inferences; not only the belief in the reported hypothesis, but also the degree of belief in reporting witnesses being reliable, as well as predictions regarding the likelihood of further corroborative reports. In the latter, we find that although lay reasoners appreciate that corroboration (or contradiction) is necessary before one can update the reliability of a reporting source (and thus avoiding a "double counting" error), we find consistent evidence of underestimation in diagnostic reasoning. Most pertinently, and novel to the literature on reasoning, this underestimation is substantially and consistently worse when the reported hypothesis base rate is rare (finding qualitative support in Experiment 1, and quantitative support in Experiments 2 and 3). We find this inverted form of base rate neglect (i.e. an under, rather than overestimation error) in the degree of support provided to a) the reported hypothesis, b) the reliability of corroborating witnesses, and c) the reliability penalisation of a contradicting minority.

Through the use of a Bayesian network formalism for the conceptualisation of coherence, the consistent deviation we find is best described in principle by a base rate neglect (Kahneman & Tversky, 1973; Tversky & Kahneman, 1981), but also fits more broadly with work on insufficient updating (e.g. Phillips & Edwards, 1966). We take pains to note that previous empirical work on the reasoning surrounding coherence has found reasonable approximations of human performance to normative (Bayesian) expectations via appropriately identifying normatively relevant factors, including witness reliability, priors, and coherence (Harris & Hahn, 2009), as well as (the relaxing of) independence constraints (Madsen, Hahn, & Pilditch, 2018). However, integrative performance in general has been shown to suffer when complexity increases, though this has typically been assessed in qualitative judgments (see e.g. Pilditch, Hahn, & Lagnado, 2018; Pilditch, Fenton & Lagnado, 2019), with deleterious performance found when witness reliabilities differ across a single integration (Phillips, Hahn, & Pilditch, 2018). Now, in the present work, we show that the failure to account for base rates in cases of corroboration leads to substantial underestimation of the multiplicative impact the two have in conjunction on diagnostic and intercausal inference. In so doing, we extend the contexts under which base rate neglect (Kahneman & Tversky, 1973; Tversky & Kahneman, 1981) is observed, simultaneously raising the issue of base rates within the theoretical context of corroborative testimony (see e.g. Harris & Hahn, 2009). Our findings confirm the theoretical predictions of Cohen (1977) that reasoners will fail to appreciate that *a priori* less likely testimony should have more impact than commonplace testimony (*ceterus paribus*) when independently corroborated as predicted by the Bayesian normative framework (Bovens & Hartmann, 2003; Pearl, 1988).

The issue of how the base rate of a reported hypothesis should influence the strength of corroborative evidence is an important one. Seldom is it appropriate, whether in applied domains (e.g. legal, forensic, or intelligence analysis) or everyday reasoning to assume that the prior hypothesis is as likely true as false (i.e., an odds ratio of 1 between hypothesis and its negation). For example, an intelligence analyst may be inclined to dismiss a claim made by a source on the ground that the terrorist in the hideout is Osama bin Laden, despite corroboration from a second source, when in fact the analyst should be *more confident* of this claim (and thus act upon it, rather than dismiss it) than a more typical claim (e.g. a high ranking officer is inside). Consequently, the implications of this work to applied domains are two-fold: first, that the base rate (or likelihood) of a given claim is attended to carefully, and second, that the independence of sources from one another in providing reports is critical (i.e. correctly identifying if witnesses are compromised, e.g. having conferred with each other, is pivotal in knowing whether the effects described herein should apply or not). Given these two elements are in place, then the decision to dismiss a claim versus believe it is heavily dependent on the degree of corroboration. In fact, given we are often interested in a target hypothesis precisely because we consider it unlikely, we are therefore often at substantial risk of the underestimations described in this paper.

AUTHOR CONTRIBUTIONS

T. D. Pilditch (TDP), D. Lagnado (DL), and S. Lagator (SL) developed the study concept. All authors contributed to the study design. Testing and data collection were performed by TDP and SL. TDP and SL performed the data analysis and interpretation with oversight from DL. TDP drafted the manuscript, and DL and SL provided critical revisions. All authors approved the final version of the manuscript for submission.

REFERENCES

Agena Ltd (2018). AgenaRisk (www.agenarisk.com)[Computer software].

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211-233.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology.* Oxford: Oxford University Press.

Bovens, L., & Hartmann, S. (2005). Why there cannot be a single probabilistic measure of coherence. *Erkenntnis*, *63*(3), 361-374.

Bovens, L., & Hartmann, S. (2006). An impossibility result for coherence rankings. *Philosophical Studies*, *128*(1), 77-91.

Cohen, L. J. (1977). *The probable and the provable*. Oxford: Clarendon.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, 5, 1-17.

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*(3), 405-425.

Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Fitelson, B. (2003). A Probabilistic theory of Coherence. *Analysis, 63,* 194-199.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological science in the public interest, 8(2)*, 53-96.

Glass, D. H. (2002, September). Coherence, explanation, and Bayesian networks. In *Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 177-182). Springer, Berlin, Heidelberg.

Glass, D. H. (2007). Coherence measures and inference to the best explanation. *Synthese*, *157*(3), 275-296.

Harris, A. J., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1366.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, 128(1), 3.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

JASP Team (2018). JASP (Version 0.9)[Computer software].

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, *4*(1), 46-63.

Lyon, D., & Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica*, 40(4), 287-298.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish

(Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 722-727). Austin, TX: Cognitive Science Society.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, 88(5), 375.

Olsson, E. J. (2002). What is the Problem of Coherence and Truth?. *The Journal of Philosophy*, *99*(5), 246-272.

Olsson, E. J. (2005). The impossibility of coherence. *Erkenntnis*, *63*(3), 387-412.

Neil, M., Fenton, N., & Nielson, L. (2000). Building large-scale Bayesian networks. *The Knowledge Engineering Review*, *15*(3), 257-284.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science*, 30(2), 250-260.

Pilditch, T.D., Hahn, U., & Lagnado, D. (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 884-889). Austin, TX: Cognitive Science Society.

Phillips, K., Hahn, U., & Pilditch, T. D. (2018). Evaluating testimony from multiple witnesses: single cue satisficing or integration? In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2244-2249). Austin, TX: Cognitive Science Society.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of experimental psychology*, *72*(3), 346.

Shogenji, T. (1999). Is coherence truth conducive? *Analysis*, *59*(264), 338-345.

Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, 109(3), 369.

Tversky, A., & Kahneman, D. (1981). *Evidential impact of base rates* (No. TR-4). Stanford Univ Ca Dept of psychology.