Title
Evaluating testimony from multiple witnesses: consistent undervaluing and selective devaluing of corroborating reports

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Phillips, Kirsty
Hahn, Ulrike
Pilditch, Toby D

Publication Date
2023

Peer reviewed

Evaluating testimony from multiple witnesses: consistent undervaluing and selective devaluing of corroborating reports

Kirsty Phillips (kphill06@student.bbk.ac.uk) & Ulrike Hahn (u.hahn@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck, University of London,
Malet Street, London, WC1E 7HX, U.K.

Toby D. Pilditch (t.pilditch@ucl.ac.uk)
University of Oxford,
Dyson Perrins Building, South Parks Road, Oxford, OX1 3QY

Abstract

This study identified a novel and robust reasoning error. Lay reasoners significantly deviate from the prediction of Bayesian inference by consistently underestimating the added probative value of corroborating testimonial reports. Most surprisingly, however, is that in certain contexts the sum of corroborating evidence is considered to be significantly less valuable than a single report. There is a selective devaluing of corroborating testimony when a highly reliable report is corroborated by less reliable, but credible, reports. This intuitive error is not explained by an inaccurate understanding of individual cues of reliability and number of reports, but specifically when it is required to integrate these both cues. Findings indicate the operation of alternative reasoning strategies, resulting in errors at individual and group level.

Keywords: Judgment; Reasoning; Decision Making; Evidence Evaluation; Corroboration; Testimony; Bayes Theorem

Introduction

"Testimony is a vital and ubiquitous source of knowledge" (Lackey, 2006, p. 432). Testimony is not only commonplace in everyday life, it is both viewed as an extremely useful and necessary reasoning processes by which beliefs are formed, and as a means to obtain knowledge from events that could not be observed directly (Adler, 2006).

However the every day use of testimony is at odds with the relative paucity of academic literature examining the use and value of testimony. Ancient Greek philosphers such as Plato and Socrates were reluctant to accept the value of testimony as a potential source of obtaining knowledge, this view persisted and was upheld by empiricists such as Locke and within contemporary epistemology (Adler, 2006; Coady, 1994). Therefore, testimony, as a source of knoweldge, has been long neglected in academic literature (Adler, 2006).

However, within the last few decades there has been increasing interest in examining the question of whether testimony can in fact give rise to knowledge, within the domains of philosophy and psychology (Coady, 1994; Lipton, 1998; Hahn, Oaksford, & Harris, 2013); including applied fields, such as forensic psychology (e.g., Winter & Greene, 2007), as well as 'core' disciplines such as cognitive (e.g., Harris & Hahn, 2009) and developmental psychology (e.g., Durfkin & Shafto, 2016). This interest also coincides with decades of psychological research demonstrating the inherent fallibility of eyewitness testimony (Loftus, 2019).

Assessing the Value of Testimony

The question therefore becomes, to what extent *can* we justifiably rely on the word of others, and *should* we revise our beliefs based on testimony alone? When testimony constitutes an argument, a proposition that speaks to values only, it is considered a fallacy to dismiss an argument based on the reliability of the source rather than the content of their testimony (Hahn et al., 2013). However, when testimony constitutes evidence, a statement that speaks to facts or a ground truth, it is necessary to assess both the reliability of the speaker and the accuracy of their statements (Hahn et al., 2013). Bayes Theorem provides a normative framework for integrating information about reliability and statement accuracy, and moreover how to optimally revise held beliefs given new evidence (Bayes, 1763). This framework has been previously used in coherence-based models of evidential reasoning, such as the Bayesian source credibility model (Bovens & Hartmann, 2003; Harris & Hahn, 2009).

Are Lay Reasoners Optimal or Irrational?

Lay reasoner performance that approximates a normative expectation has been termed evidence of the 'statistical man'; a body of research has found that individuals are influenced by the appropriate variables and in the correct direction (Peterson & Beach, 1967). The most recent evolution of this theory (due to increased adoption of Bayes Theorem within cognitive psychology as a normative framework) is the Bayesian brain hypothesis (Williams, 2021). Broadly individuals successfully navigate an inherently uncertain world (Chater et al., 2011) and research evidence from various complex domains (e.g., Knill & Pouget, 2004) supports the proposition that cognitive processes are approximately Bayesian (Knill & Pouget, 2004).

However, the veracity of the Bayesian brain hypothesis continues to be debated (Williams, 2021). There is a competing body of literature, similarly, using a Bayesian framework as a benchmark for optimal performance, that finds systematic deviations from the Bayesian prediction

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

(e.g., Tversky & Kahneman, 1974). This deviation is termed error, bias, or irrationality (Mellers, Schwartz, & Cooke, 1998; Tversky & Kahneman 1974). This deviation is thought to be the result of intuitive, 'system 1', cognitive processes (Kahneman, 2003), which are considered to be more error prone and utilise alternative cognitive processes, such as 'satisficing' (Simon, 1956) or the use of simpler heuristic strategies (Gigerenzer et al., 1999).

Assessing the Value of Corroborating Testimony

Testimony does not often occur in isolation. In the case of multiple witnesses, it is necessary to integrate not only knowledge of reliability but also the number of witnesses and any potential dependency between them (see Bovens & Hartmann, 2003) in order to maximise the accuracy of beliefs (Pettigrew, 2016). There are different forms of corroborative witness testimony (see Redmayne, 2000). One form is simply where two or more witnesses independently report that the same event occurred (jointly supporting or opposing the hypothesis under consideration); this is termed "same fact corroboration" (Redmayne, 2000) or "corroboratively redundant testimony" (Schum & Martin, 1982). Bayes Theorem (1763) provides a normative framework for integrating multiple pieces of evidence and how to optimally revise held beliefs given evidence from multiple sources (see Schum & Martin, 1982). Corroborating testimony by definition involves more than one witness, therefore this represents a complex evidence structure (Schum & Martin, 1982); a reasoner must assess the probative value of individual elements and how these elements interact to determine their combined value.

It is this reasoning process which poses an interesting empirical question, do lay reasoners intuituively approximate optimal performance, as predicted by Bayesian inference, when integrating informational cues concerning reliability and number of independent testimonies? Or do lay reasoners revert to simpler, heuristic "operating rules", which rely on processes involving single cues, resulting in intuitions that are contrary to Bayesian inference? Previous research, such as behavioural models of persuasion (Chaiken, 1980; Petty & Cacioppo, 1986), have found that individuals seemingly focus on single salient cue.

Phillips, Hahn and Pilditch (2018) identified a novel intuitive reasoning in evaluating the value of corroborating testimony, concluding that most pariticpants engaged in some form of satisficing when evaluating hypothetical combinations of eyewitness reports. In this prior study, participants were presented with a simplified scenario involving multiple witness reports. Participants were asked to imagine there are the manager of a business investigating missing petty cash, with the hypothesis to evaluate is whether the cash is missing due to an error or due to theft. There are five potential witnesses, one employee, Chris, was much more reliable (at 95% hit rate, described as reporting evidence of wrong doing), compared to the other four (Alan, Brad, David, and Edward, with a 15% hit rate). It was explained that employees may hesitate to give a report about events of wrongdoing and in all other instances they will remain silent. It was also stated that the false positive rate was low consistent across all employee reports; "You do not believe that any of your employees would lie, (i.e., claim that cash was stolen when it was not), so we can assume the chance of this is low for all (say, 10%)". Participants were presented with varying combinations of witness reports (scenarios shown in Table 1) and asked to rank them ("which of the following scenarios would convince you that cash was in fact stolen?") from most to least convincing. This experiement scenario was simplified by design. Firstly, the difference in reliability was signified by hit rate alone, so participants did not need to assess mutliple factors to determine reliability. Secondly, it was not required for participants to consider prior or post probabilities to complete the task, they were asked simply to consider evidential weight of combinations, which remains constant irrespective of prior belief.

Table 1: Witness combination scenarios presented in Phillips et al. (2018), shown are the witness combinations ranked by likelihood ratio, calculated using Bayesian inference.

Rank	Witness Combination Scenarios	Likelihood Ratio
1	"Chris & Alan" (C&A)	14.25
2	"Chris" (C)	9.5
3	"Alan, Brad, David & Edward" (A,B,D&E)	5.0625
4	"Brad & David" (B&D)	2.25
5	"Edward" (E)	1.5

Shown above, in Table 1, is the correct rank order and probative value (termed likelihood ratio) of each of the witness combinations. Adopting terminology used by Schum (1981), the "hit rate" (alternatively termed true positive rate) corresponds to the probability of receiving a confirmatory report if the hypothesis is in fact true ($P(e|H)$) whereas the "false positive rate" corresponds to the probability of receiving a confirmatory report if the hypothesis is acutally false ($P(e|\neg H)$). The likelihood ratio, or diagnostic impact of individual and collective evidence, is calculated by dividing the hit rate by the false positive rate ($P(e|H)/P(e|\neg H)$). In cases of multiple pieces of evidence, the respective hit rates and false positive rates are multiplied and the sum hit rate and false positive rate are divided (e.g., for the C & A scenario the likelihood ratio is calculated by $0.95*0.15/0.1*0.1 = 0.1425/0.01 = 14.25$). A likelihood ratio of greater than 1 should always increase beliefs in a hypothesis, but less than 1 should always decrease beliefs in a hypothesis and 0 is deemed non-diagnostic or irrelevant to the hypothesis in question. Therefore, all the witnesses in this study were credible, as all had a likelihood ratio of greater than 1, but some are more reliable than others, meaning their evidence is closer aligned to ground truths, and is therefore more diagnostic.

The combinations were chosen to test whether participants can accurately combine number of reports given

equal reliability (i.e., A, B, D & E > B & D > E), accurately determine differences given unequal reliability (i.e., C > E) and combine both cues of number and reliability (i.e., C & A > C).

Phillips, Hahn and Pilditch (2018) determined that participants generally failed to adhere to the correct rankings indicated by a normative (Bayesian) standard, with only 8.3% ranking all scenrios in the correct order. Whilst the majority did accurately combine number of reports given equal reliability (B & D as rank 4 and E as rank 5) and most accurately determined differences given unequal reliability (most ranked E as the least convincing scenario), participants seemingly had difficulty determining which of the the top three scenarios (1-3 in Table 1) were most convincing; at a group level, there was no difference in how the top three options are ranked. This indicated that subgroups adopted alternative satisficing strategies (relying on individual cues of reliaiblity, C, or number, A, B, D & E), as no tie ranks were permitted in the task. Most suprisingly, only the minority (13.33%) accurately combined both number and reliability cues (ranking C & A above C), thereby adopting intuitive reasoning strategies which approximated the predications of Bayesian inference.

Within the current study we are interested in uncovering and confirming the specific inability of lay reasoners to accurately incorporate additional corroborating evidence from sources of lower reliability. We will further explore potential explanations (i.e., inaccurate evaluation of individual informational cues concerning reliability and report number). We will also test the robustness of this error by determining if the error persists when further corroborative evidence is given, to see if this erroneous intution results in further deviation from the conclusions of Bayesian inference.

Present Study

The aim of this study is to further examine the reasoning error identified in Phillips et al. (2018). This study will adopt the same hypothetical scenario as in the original experiment. Participants were presented with the same simple reasoning task; they were the manager of a business in which petty cash had gone missing, the target hypothesis was to evaluate whether there had been a theft. As within the previous study, there are five employees who are potential witnesses. One of the witnesses, "Chris", was stated to be much more reliable (at 95% hit rate), compared to the other four witnesses ("Alan", "Brad", "David", and "Edward"; each at 15% hit rate); described as "reports wrong-doing on occasions when wrong-doing has actually occurred". The false positive rate was low at 10% and consistent across all employee reports, so that as in with the previous study variation in reliability was defined by hit rate alone, and described as "to claim that cash was stolen when it was not". Within the scenario participants were asked to consider "the probability of the cash being stolen, given the reports in each of the following scenarios", after having collected "statements from each of these employees separately and each report that the cash was stolen". The initial prior belief in probability of theft was stated to be 50%. Table 2 shows the likelihood ratio (probative value independent of prior belief) and correct posterior probabilities, calculated via Bayes' Theorem, for each witness combination used in this study. According to Bayesian inference, prior odds (in this study a 50% prior probability was used, equivalent to 1 or equal odds) is multiplied by the likelihood ratio (termed post odds), and then divided by post odds plus 1 to obtain the posterior probability ($P(Theft|E_{1\text{-}N})$). For example, for the C&A scenario the post odds is 1*14.25 = 14.25 and the posterior probability is 14.25 /(14.25 +1) = 0.9344, or 93.44%.

Table 2: Witness combination scenarios presented in the current study. Using Bayesian inference, shown are the witness combinations ranked by likelihood ratio and resulting posterior probabilities for the target hypothesis, given the reports.

Rank	Witness Combination Scenarios	Likelihood Ratio	$P(Theft\|E_{1\text{-}N})$
1	"Chris, Alan & Edward" (C,A&E)	21.375	95.53%
2	"Chris & Alan" (C&A)	14.25	93.44%
3	"Chris" (C)	9.5	90.48%
4	"Alan, Brad, David & Edward" (A,B,D& E)	5.0625	83.51%
5	"Brad & David" (B&D)	2.25	69.23%
6	"Edward" (E)	1.5	60%

Two key changes were made to the methodology, to further explore potential explanations and test the robustness of this error. Firstly, instead of five, six witness combinations were presented. In addition to the five combinations used in the original study, an additional corroborating scenario was added; where a single high reliability report was corroborated by a two less reliable reports ("Chris, Alan and Edward"). This scenario was added to examine whether the identified reasoning error persists with additional corroborating reports. Secondly, rather than rank combinations, for each of the six combinations of witness reports, participants were asked to provide an estimate (of the likelihood that the cash had been stolen, given the stated combination of witness reports). Estimates were obtained using a sliding scale (0-100), with the slider starting at 50% (the initial, prior, belief in theft).

The objective of this study is to further elucidate the findings from the previous study (Phillips et al., 2018). Firstly, to determine if, like the previous study, estimates in significantly deviate from Bayesian inference. Secondly to replicate the reasoning error identified in the previous study; to investigate if estimates given for C&A are significantly less than C. In addition, to rule out task misunderstanding or absolute inability to aggregate multiple reports, estimates for B&D will be compared to estimates for E. Thirdly, to

investigate the "C&A" error further by investigating if the error persists with additional corroborating reports (from C, A & E).

Therefore, the hypotheses can be summarized as follows:

H1. Participant estimates of the six witness combination scenarios will significantly deviate from the predictions of Bayesian inference.

H2. Participants estimates of C & A will be significantly less than C. In addition, estimates of B & D will also be tested to see if estimates are significantly less than E.

H3. Participants estimates of C, A & E will be significantly less than estimates for both C & A and C.

Methods

Participants 60 (30 female) US participants were recruited and participated online through the MTurk platform. Among the participants, 33 had been educated to the level of Bachelor's Degree or above. The mean age of participants was 36.99 (*SD* = 12.09). Informed consent was obtained, and all participants were appropriately compensated for their time.

Procedure and Materials All participants completed the survey, conducted using the Qualtrics platform. The survey consisted of 11 questions in total: Qs 1-3 obtained informed consent; Qs 4-6 obtained demographic information (age, gender, and education level); Q7 obtained an MTurk ID for reimbursement; Qs 8 & 9 presented the scenario and obtained participants' estimates; Qs 10 & 11 obtained explanatory text and confidence ratings.

Analysis

The analysis is split into four parts. First, participant estimates of likelihood of theft are compared to the predictions of Bayesian inference across all six witness combinations. Second, participant estimates of likelihood of theft given two corroborative reports (C & A and B & D) are compared to estimates from a single report (C and E respectively). Third, participant estimates of likelihood of theft given three corroborative reports (C, A & E) are compared to estimates from two corroborative reports (C & A) a single report (C). Finally, plots are used to explore to indicative strategies of sub-groups of participants.

JASP (Version 0.17) statistics program software was used to conduct analysis. Shapiro Wilk tests found that estimates across all witness combinations violated the assumption of normality: C, A & E, $W(59)=.788$, $p<.001$; C & A, $W(59)=.826$, $p<.001$; C, $W(59)=.662$, $p<.001$; A, B, D & E, $W(59)=.924$, $p=.001$; B & D, $W(59)=.958$, $p=.037$; and, E, $W(59)=.945$, $p=.009$. Therefore, the following analyses were conducted using non-parametric tests.

Descriptive Findings Obtained participant estimates for each witness combination are shown in Figure 1, and

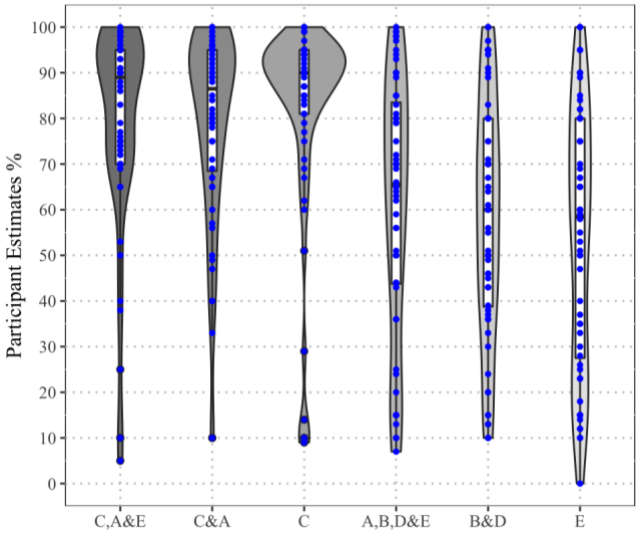

Figure 1. Violin plots with box plots of estimates across all witness combination scenarios presented in this study (N=60). Individual estimates are shown by the blue dots.

median estimates are shown in Table 2. Overall, estimates are highest for C (a single high reliability report). Most participants (34, 56.67%) do not increase their estimates when C's report is corroborated by A's report (less reliable but still credible). Similarly, most participants (34, 56.67%) do not increase their estimates when C & A's report is supplemented by E's report (E's report is of equal value to A's individual report). Only 30% (N=18) correctly provided estimates to indicate that the likelihood of theft should increase in both these instances (i.e., that C, A & E > C & A and C & A > C). However, when reports are of equal (lower) reliability participants are less likely to make this error. Most participants (34, 56.67%) increase their estimates when B & D's reports are in corroboration, compared to estimates for E's single report (when B, D & E's report are of equal reliability). Unlike the previous study, 70% (N=42) of participants correctly provided estimates to indicate that four corroborating reports of lower reliability (A, B, D & E) are less probative than a single report from a highly reliable witness (C).

Comparison to Bayesian Inference (H1) Six two-tailed Wilcoxon Signed-Rank Tests were conducted; estimates across the six witness combination scenarios were tested against values predicted by Bayesian inference, to determine if participant estimates significantly deviate from optimal. The results of these analysis, shown in Table 2, find that only the corroborative scenarios significantly deviate from the normative prescription. In each instance the value of corroborative reports is less than the normative prescription, meaning corroborative reports are consistently undervalued.

Table 2: Outcome of analyses using JASP (Version 0.17). Included are test scenarios, test values, witness combinations medians, outcome of Wilcoxon Signed-Rank Tests, Effect sizes (including 95% CI) and Vovk-Sellke Maximum *p*-Ratio (maximum possible odds in favour of H1 over H0, when $p \leq .37$) (Sellke, Bayarri, & Berger, 2001).

Scenarios	Median	Test value	Wilcoxon Signed-Rank Tests		Effect Size	95% CI Effect Size		VS-MPR
			W	*p*		Lower	Upper	
H1: Comparison to Bayesian Inference								
C, A & E	89	95.53	187	<.001	-.769	-.880	-.662	269019.234
C & A	86.5	93.44	328	<.001	-.624	-.782	-.438	2133.802
C	90	90.48	656	=.057	-.283	-.524	-3.132×10^{-5}	2.261
A, B, D & E	65.5	83.51	252	<.001	-.725	-.836	-.555	24961.149
B & D	60	69.23	552	=.008	-.397	-.611	-.128	9.917
E	58.5	60	623.5	=.108	-.246	-.500	-.048	1.534
H2: Two corroborating reports are devalued								
C & A < C	86.5	90	430	<.001	-.497	-	-.295	97.829
B & D < E	60	58.5	908	=.481	-.008	-	-.232	1.000
H3: Further corroborating reports are devalued								
C, A & E < C	89	90	347.5	<.001	-.496	-	-.279	66.467
C, A & E < C & A	89	86.5	663.5	=.032	-.275	-	-.038	3.326

Two corroborating reports are devalued (H2) Two one-tailed Wilcoxon Signed-Rank Tests were conducted; estimates for C & A were tested against estimates for C and estimates for B & D were tested against estimates for E, to determine if two corroborative reports are estimated to have significantly less value than a single report. The results of this analysis, shown in Table 2, find that estimates for C & A are significantly less than estimates for C. However, estimates for B & D are not significantly less than estimates for E.

Further corroborating reports are devalued (H3) Two one-tailed Wilcoxon Signed-Rank Tests were conducted; estimates for C, A & E were tested against estimates for C & A and C, to determine if additional less reliable reports are believed to further devalue a single high reliability report. The results of this analysis, shown in Table 2, find that estimates for C, A & E are significantly less than estimates for both C & A and C.

Exploratory Analysis The results of these significant findings, in relation to H2 & H3, are further explored using plots (shown in Figure 2), to explore whether subgroups do adopt alternative satisficing strategies as hypothesised by Phillips et al. (2018). Estimates were divided according to whether estimates increased (indicative of optimal reasoning strategies), or were equal or decreased (indicative of alternative reasoning strategies), relative to other estimates: C, A & E compared to C & A; C, A & E compared to C; and, C & A compared to C. Most participants provided estimates which correspond to non-optimal judgments (i.e., estimates were not increased). Additionally, participants were equally likely to increase or decrease estimates for C, A & E compared to C. The plots show that the median of both the 'increase' groups (41.67%-43.33%) and 'equal' groups (16.67-25%) approximate the predictions of Bayesian inference; although there is much greater variation in 'equal' groups. Importantly, as shown by the variation in individual estimates in the 'increase' and 'equal' group, participants are not simply anchoring on the given hit rate value of C (95%). It is the 'decrease' group (31.67%-41.67%) only which appear to deviate from the predictions of Bayesian inference, with given estimates being substantially lower than optimal.

Discussion

This paper sought to investigate the capacity of lay reasoners to integrate accurately both the reliability and number of independent testimonies, to further explore the findings of Phillips et al. (2018). This study was able to demonstrate that participants have difficulty in appreciating the added value of multiple reports. Participant estimates in relation to corroborating witness reports signifcantly deviated from the predictions of Bayesian inference; in all instances the estimates were conservative and therefore the added value of corrobroative reports was consistently underestimated. However, conservatism does not explain the specific error first identified by Phillips et al. (2018) and the robust 'devaluing error', uniquely demonstrated within this study. Participants have a specific inability to accurately integrate cues of both reliability and number of reports, resulting in the devaluing of corroborating reports, when the combination of witness reports includes a single report of high reliablity and supplementary reports of lower (but still credible) reliability.

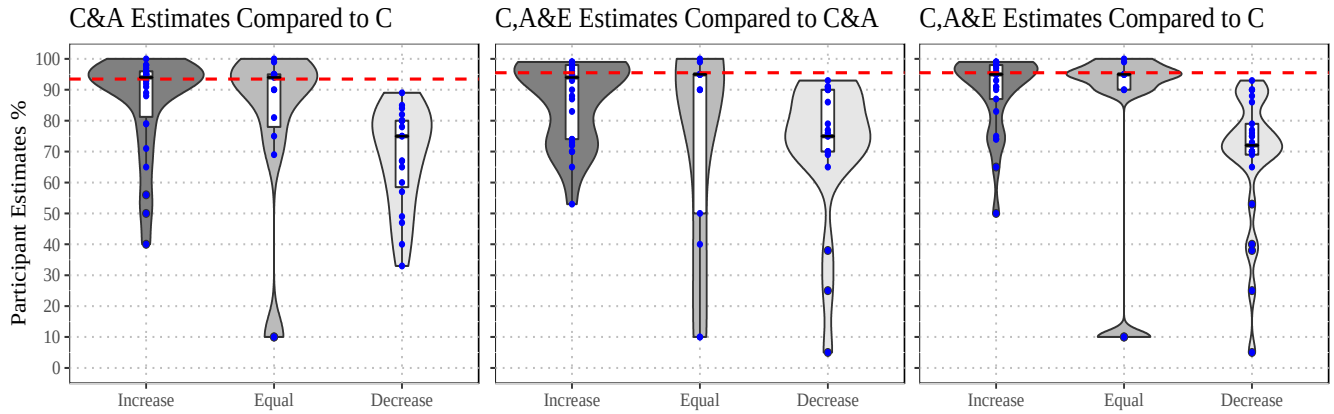

Figure 2. Violin plots with box plots to compare possible alternative reasoning strategies adopted (to increase, decrease estimates or keep estimates equal) when a single high reliability report (C) is corroborated by one less reliable report (C & A) and a second less reliable report (C, A & E). Left plot: Increase N=26, Equal N=15, Decrease N=19. Middle plot: Increase N=26, Equal N=13, Decrease N=21. Right plot: Increase N=25, Equal N=10, Decrease N=25. The Bayesian prediction is shown by the red dotted line, individual estimates are shown by the blue dots.

This study has eliminated the possibility that this errror results from a misunderstanding or misinterpretation of informational cues concerning reliability. Within this study, it was determined that participants estimations of the influence of single testimonal reports, of differing levels of reliability, correspond to the predictions of Bayesian inference. Importantly, particpants were able to accurately use reliablity cues to correctly assess that C's single highly report was more probative than E's single less reliable report, and in both instances beliefs were increased. Furthermore, this study was able to eliminate the possibility that this error results from a general inability to incorporate additional reports. In scenarios in which the reliability is held constant, even though participants underestimated the value of additional reports, participants broadly provide estimates in the correct order; that a group of four reports increases the likelihood compared to two reports, and that two reports increases the likelihood compared to a single report. Therefore, broadly, participants are able to accurately use cues about number of reports, when reliabiltiy is held constant, to correctly assess that corroboration increases the likelihood of the hypothesis.

Importantly, not only has this study replicated the specific reasoning error identified in the original study and been able to eliminate explanations relating to individual informational cues, this study has further shown that this specific reasoning error persists when further additional less reliable reports are added; resulting in further devaluing of corroborating reports. Exploratory analysis indicates that some participants are engaging in intuitive strategies which approximate the predictions of Bayesian inference. However, most participants are adopting intuitive strategies that would lead to conclusions that are in contradiction to the predictions of Bayesian inference. Alternative strategies that would lead to a decrease in estimates could suggest evidence of averaging strategies (e.g., Lopes, 1985) or 'dilution effects' (Madsen, Hahn, & Vorms, 2017). Alternative strategies that would lead to equal estimates (therefore no impact of additional corroborating reports) correspond to the predictions of the MAXMIN rule (see Walton, 1992, 2007). However, these strategies may only offer a partial explanation, as these alternative strategies were not applied across all scenarios, but only when it is necessary to integrate both cues of reliability and number of reports. As this study demonstrates, participants are able to recognise that all reports in this scenario are diagnostic and hold probative value, by accurately incorporating reports of low reliability (by adjusting estimates upward) when all corroborating reports are of equal reliability. Therefore, the predictions of averaging strategies, dilution effects and the MAXMIN rule do not hold in these circumstances. Future work could further explore the use of qualitative judgements when considering corroborating testimony, to determine if these fallacious judgements are overtly made when considering the impact of additional reports.

Conclusions

Overall, this study concurs with previous findings that lay reasoners do not integrate corroborative testimonies in the manner expected by normative, Bayesian standards. Lay reasoners consistently underestimate that added value of corroborative reports. Yet, there is evidence that participants understand individual cues of reliability and number of reports. A specific and robust reasoning error has been identified, which is obviously problematic as, normatively speaking, additional independent reports will always add value as long as they are somewhat reliable and diagnostic. However, as this study uniquely demonstrates that most lay reasoners do not believe that corroborating reports of lower reliability add any value and in fact many believe the initial evidence is devalued, or 'tainted', in the light of additional lower reliability reports. This results in error at the individual and group level. Further work is needed to understand these erroneous intuitions.

Acknowledgments

This research is in part based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under Contract [2017-16122000003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Authors KP and UH formulated the study design, author KP conducted data collection and analysis, and KP, TP, and UH wrote the manuscript.

References

Adler, J. (2006). Epistemological problems of testimony. *The Stanford Encyclopedia of Philosophy*. Stanford University.

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological bulletin, 82(2)*, 261-277.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions, 53*, 370-418.

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

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology, 39(5)*, 752-766.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2011). Inductive logic and empirical psychology. In D. M. Gabbay, S. Hartmann, & J. Woods, *Inductive Logic. Handbook of the History of Logic 10.* Elsevier.

Coady, C. A. (1994). *Testimony: A Philosophical Study.* Oxford: Oxford Academic.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart* (pp. 75-95). Oxford University Press.

Hahn, U., Oaksford, M., & Harris, A. J. (2013). Testimony and argument: A Bayesian perspective. In F. Zenker, *Bayesian argumentation: The practical side of probability* (pp. 15-38). Netherlands: Springer.

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 –1373.

JASP Team. (2003). JASP (Version 0.17)[Computer software].

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449-1475.

Knill, D. C., & Pouget, A. (2004). The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation. *Trends in Neurosciences, 27*(12), 712-719.

Lackey, J. (2006). Knowing from Testimony. *Philosophy Compass, 1*(5), 432–448, DOI:10.1111/j.1747-9991.2006.00035.x.

Lipton, P. (1998). The epistemology of testimony. *Studies in History and Philosophy of Science Part A, 29 (1)*, 1-31.

Loftus, E. F. (2019). Eyewitness testimony. *Applied Cognitive Psychology, 33(4)*, 498-503.

Madsen, J. K., Hahn, U., & Vorms, M. (2017). The dilution effect: Conversational basis and witness reliability. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Ed.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* CogSci.

Mellers, B. A., Schwartz, A., & Cooke, A. D. (1998). Judgment and decision making. *Annual review of psychology, 49(1)*, 447-477.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological bulletin, 68(1)*, 29-46.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology, 19*, 123-205.

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 (Ed.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2244-2249). Proceedings of the 40th Annual Conference of the Cognitive Science Society .

Redmayne, M. (2000). A corroboration approach to recovered memories of sexual abuse: A note of caution. *Law Quarterly Review, 116(Jan)*, 147-155.

Schum, D. A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance, 27(2)*, 153-196.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review, 17(1)*, 105-152.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician, 55*(1), 62-71.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review, 63(2)*, 129-138.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185(4157)*, 1124-1131.

Walton, D. (1992). Rules for plausible reasoning. *Informal Logic, 14(1)*, 33-51.

Walton, D. (2007). *Witness Testimony Evidence: Argumentation and the Law.* Cambridge University Press.

Williams, D. (2021). Epistemic irrationality in the Bayesian brain. *The British Journal for the Philosophy of Science, 72*(4), 913-938.

Winter, R. J., & Greene, E. (2007). Juror decision-making. In F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandowsky, & T. J. Perfect, *Handbook of applied cognition, 2* (pp. 739-762). Chichester, England: John Wiley & Sons, Ltd.