Evaluating testimony from multiple witnesses: exploring qualitative intuitions

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)
Department of Experimental Psychology, University College 26 Bedford Way, London, WC1H 0AP, U.K.

Abstract

This study further explored a novel reasoning error. When faced with evidence from multiple sources, a substantial number of lay reasoners inaccurately integrate cues of reliability and report number. Particularly when further reports are less reliable than initial (highly reliable) reports. When evaluating the added value of supplementary corroborative reports, we find that, in most instances, participants are equally likely to provide correct or incorrect qualitative judgements. When using a sequential presentation and explicitly prompting participants to consider the impact of additional credible evidence, 36.7%-45% indicate that their beliefs should remain the same and 10% or less indicate that their beliefs should decrease. Only a third correctly believed that in each instance of corroborating evidence the likelihood of the target hypothesis should increase. Qualitative judgements also significantly impacted the accuracy of belief estimates; deviations from normative, Bayesian, predictions at the group level are explained by sub-groups with incorrect qualitative intuitions.

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); it serves as an extremely useful and necessary means to obtain knowledge about events that cannot be observed directly (Adler, 2006). However, there has been a long-held scepticism as to the epistemological value of testimony, simply put, 'can the word of others truly give rise to knowledge?' (Adler, 2006; Coady, 1994). Recently, there has been renewed interest in examining this question within academic literature (e.g. Harris & Hahn, 2009; Hahn, Oaksford, & Harris, 2013; Durfkin & Shafto, 2016). Concurrently, a body of psychological literature has explored the (un)reliability of eyewitness testimony (see, Loftus, 2019). Testimony is widely accepted to be inherently uncertain, therefore, the extent to which testimony gives rise to knowledge is a probabilistic problem. Bayes' Theorem provides a normative framework for optimally integrating information about the content of an eyewitness' report, the reliability of this report, and how to revise held beliefs given this evidence (Bayes, 1763). Approximation of normative frameworks is termed the 'statistical man' (Peterson & Beach, 1967) or more recently the Bayesian brain hypothesis (Williams, 2021); supported by findings that individuals do successfully navigate real-world uncertainty (Chater et al., 2011), are influenced in the correct direction by the appropriate variables (Peterson & Beach, 1967) and cognitive processes approximate Bayesian predictions (Knill & Pouget, 2004). However, deviation from normative frameworks is considered an error, bias, or irrationality (Mellers, Schwartz, & Cooke, 1998; Tversky & Kahneman 1974) and evidence of intuitive ('system 1') processes, such as 'satisficing' (Simon, 1956) or heuristics (Gigerenzer & Goldstein, 1999).

Evaluating Corroborating Testimony

Additional testimonial reports increase the complexity of the probabilistic problem; to determine diagnosticity of evidence it becomes necessary to integrate report content, report reliability and evidence structure (Bovens & Hartmann, 2003; Schum & Martin, 1982). An example of evidence involving multiple testimonial reports is corroborating testimony, of which there are different forms (see Redmayne, 2000). The example we will explore in this paper is termed 'same fact corroboration' (Redmayne, 2000) or 'corroboratively redundant testimony' (Schum & Martin, 1982); where two or more witnesses independently agree, by supporting (or opposing) the hypothesis under consideration. Bayes' Theorem also provides a normative framework for determining the combined value of corroborative testimony and how to accurately revise held beliefs given this combined evidence (see Schum & Martin, 1982). Briefly, diagnosticity is quantified by the likelihood ratio (LR); the 'hit rate' (sensitivity) divided by the 'false positive' (1-specificity). The combined value (combined LR) of corroborating testimony is simply individual LRs multiplied. Prior odds multiplied with the combined LR derives the posterior odds. Posterior odds divided by posterior odds plus one derives the post probability. A normative, Bayesian, evaluation predicts that a LR of >1 should always increase beliefs and add to the value of combined evidence. Given the tension between the inherent complexity of the problem and the fact that lay reasoners often evaluate testimony from multiple sources in both formal and informal settings, this poses an empirical question in need of exploration, i.e., when evaluating corroborating testimony do lay reasoners intuitively

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

approximate optimal, Bayesian, predictions or do lay reasoners intuitions reveal errors that lead them astray?

Errors in Evaluating Corroborating Testimony

A novel and robust reasoning error in relation to corroborating eyewitness testimony has been identified by Phillips, Hahn and Pilditch (2018; 2023); a specific intuitive error that causes lay reasoners to inaccurately aggregate corroborating evidence, particularly when additional reports are from sources of lower reliability. Across both studies participants were asked to consider the same hypothetical scenario; they were asked to imagine they were the manager of a business in which petty cash had gone missing, the target hypothesis being whether a theft had occurred. Five employees could have been potential witnesses, yet these witnesses differed in their reliability. The false positive rate was described as "to claim that cash was stolen when it was not"; this was kept constant and low (stated to be 10%) across all employee reports, so that variation in reliability was defined by hit rate alone. The hit rate was described as "reports wrong-doing on occasions when wrong-doing has actually occurred" and stated to be 15% for four of the witnesses ("Alan", "Brad", "David", and "Edward") and 95% for one witness ("Chris"). Therefore, reports from Chris (LR=9.5) should be considered much more reliable and therefore more convincing than the other four witnesses (LR=1.5). Importantly, all reports are diagnostic, and should be considered credible as their hit rate is greater than their false positive (LR >1), therefore all reports should always increase beliefs and add value to combined evidence.

In the first study (Phillips et al., 2018), participants were asked to rank five hypothetical combinations of witness reports from most to least convincing. The combinations presented (in the correct order) were Chris and Alan (C&A), Chris (C), Alan, Brad, David and Edward (A,B,D&E), Brad and David (B&D) and Edward (E). Only 5 participants (8.3%) proceeded to rank all options correctly, demonstrating intuitive reasoning strategies which approximated optimal Bayesian predictions. However, the most surprising finding was that only 8 participants (13.3%) correctly recognised that the combination of Chris's report when corroborated by Alan (C&A) was more convincing than Chris's (C) single report. Participants instead showed a preference for either a single report of high reliability (C) or the highest number of reports from witnesses of lower but equal reliability (A,B,D&E). In contrast, participants had no difficulty recognising that when reliability is held constant across reports four reports are more convincing than two, and two reports are more convincing than one (A,B,D&E>B&D>E). Therefore, findings indicated that participants were engaging in some form of satisficing. Meaning a preference was shown for single cues of either reliability or number of reports; the majority did not utilise both cues accurately.

In the second study (Phillips et al., 2023) rather than rank the value of these combinations, participants were asked to estimate the posterior probability given the hypothetical combinations of eyewitness reports and a 50% prior probability of theft (prior odds of 1). An additional combination of Chris, Alan and Edward (C,A&E) was added to further test the identified reasoning error; to determine if errors persist even when further corroborative evidence is given. Findings demonstrated that only estimates for corroborating witness scenarios significantly differed from Bayesian predictions, whereas estimates for single witness scenarios did not. Participant estimates in response to corroborating reports only were conservative, meaning the added value of corroborative reports was consistently underestimated. Findings further demonstrated that, most surprisingly, in certain scenarios the combined corroborating evidence is considered significantly less convincing than a single initial report. But only when corroborating reports are less reliable than an initial highly reliable report. It was found that C,A&E was estimated to be significantly less than both C&A and C, and C&A was significantly less than C. In contrast, estimates for B&D were not significantly less than estimates for E. Therefore, this error cannot be explained by conservatism (Ajzen & Fishbein, 1975) or an inaccurate understanding of individual cues of reliability and number of reports. A specific intuitive error arises when it is necessary to integrate both cues. An exploratory analysis was conducted to determine proportions of qualitative response patterns, whether estimates increased, decreased, or were not adjusted (were equal). When looking at C&A vs. C, C,A&E vs. C&A, and C,A&E vs. C, patterns of response proportions were similar. The majority did not respond in the predicted direction, only 43.3%-41.7% increased their beliefs; 41.7%-31.7% decreased their beliefs and 25%-16.7% did not adjust their beliefs. Participants were equally as likely to increase or decrease estimates for C,A&E compared to C. It was determined that further study was needed to explore these inferred qualitative intuitions.

Present Study

The aim of this study is to further examine the reasoning errors identified in Phillips et al. (2018; 2023); to further understand the specific error that arises when additional corroborating evidence from sources of lower reliability are inaccurately incorporated. Previous studies eliminated other potential explanations, such as conservatism or an inaccurate understanding of singular informational cues concerning reliability and report number. Within this study we will adopt the same hypothetical scenario as in the original experiments, but adopt an alternative methodology. We will measure qualitative intuitions by asking participants to make qualitative judgments concerning the impact of corroborating evidence on their beliefs (increase, decrease or no impact) before obtaining belief estimates. This format necessitates another change, whereas in the two previous studies hypothetical combinations were presented simultaneously, in this study combinations will need to be presented sequentially; to measure qualitative judgements and belief estimates at each stage when new evidence is

Table 1: Bayesian predictions shown for the witness combinations in relation to rank, likelihood ratio (LR) and posterior probabilities (Post *P*).

Witness Combinations	Rank	LR	Post *P*
Scenario 1			
Chris	3	9.5	90.48%
Chris & Alan	2	14.25	93.44%
Chris, Alan & Brad	1	21.375	95.53%
Scenario 2			
Edward	6	1.5	60%
Edward & David	5	2.25	69.23%
Edward, David & Brad	4	3.375	77.14%

introduced. The robustness of this error will also be further tested using this methodology, importantly, at each stage participants will be explicitly prompted to consider the impact of additional evidence on their beliefs.

Within this study two alternative scenarios will be compared; initial reports are either from a highly reliable witness ("Chris") or from a less reliable but still credible witness ("Edward"). Presentation order of the two scenarios will be randomised across participants. In both scenarios, initial reports are supplemented by two reports from witnesses whose reliability matches Edward's, in two stages. Additional reports are held equal across the two scenarios, so the impact can be compared. After the first initial report estimates will be obtained using a sliding scale (0-100), starting initially at the prior probability (50%). As before, the target hypothesis to consider is whether cash has been stolen, given the eyewitness reports. At each stage when new evidence is introduced, participants will be asked to select a qualitative judgement: "It is now more likely the cash was stolen", "The probability the cash was stolen remains the same" or "It is now less likely the cash was stolen". Estimates will then be obtained using a sliding scale (0-100), starting at their previous answer.

Table 1 shows the optimal, Bayesian, predictions for the six witness combinations used in this study. Including the likelihood ratio (diagnosticity independent of prior belief) and posterior probabilities (updated beliefs given the evidence); posterior probabilities are the benchmarks we will use to determine belief accuracy. Therefore, like Phillips et al. (2023), six witness combinations will be presented and include C,A&B (when an initial highly reliable report is corroborated by two less reliable reports). However, unlike previous studies, the introduction of individual reports can be directly compared across the two scenarios.

With this new methodology we will be able to assess qualitative judgements, to see if there is a clear preference at each stage. Based on the findings from Phillips et al. (2023) we can predict that correct and incorrect qualitative judgements will be selected at approximately equal rates. We will also test the impact of these qualitative judgements on belief accuracy. Based on the findings from Phillips et al. (2023) we can also predict that only those with incorrect qualitative judgements will have inaccurate beliefs, and that this explains the errors observed at group level. Therefore, the alternative hypotheses can be summarized as follows:

H1. There will be a significant difference in selection of qualitative judgements, with a significant preference for the correct qualitative judgement.

H2. Participant estimates will significantly deviate from Bayesian predictions, only when an incorrect qualitative judgement is made.

Methods

Participants. 60 (24 female) US participants were recruited and participated online through the MTurk platform. Among the participants, 28 had been educated to the level of Bachelor's Degree or above. The mean age of participants was 33.77 (*SD*=9.68). Informed consent was obtained, and all participants were appropriately compensated for their time.

Procedure and Materials. All participants completed the survey, conducted using Qualtrics. The survey consisted of 23 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 8-13 and Qs 16-21 presented the task and obtained participants' responses for scenario 1 and scenario 2 respectively; and, Qs 14-15 and Qs 22-23 obtained explanatory text and confidence ratings for scenario 1 and scenario 2 respectively.

Analysis

JASP (Version 0.17) statistics program software was used to conduct statistical analysis. Statistical analysis is split into three parts. First, proportion of correct and incorrect qualitative judgements are compared (H1). Secondly, the accuracy of belief estimates are assessed according to correct and incorrect qualitative judgements (H2). Shapiro Wilk tests found that all except one of the estimates violated the assumption of normality. Correct: C&A, $W(27)=.628$, $p<.001$; C,A&B, $W(31)=.761$, $p<.001$; E&D, $W(33)=.935$, $p=.045$; and, E,D&B, $W(30)=.908$, $p=.011$. Incorrect: C&A, $W(31)=.788$, $p<.001$; C,A&B, $W(27)=.711$, $p<.001$; E&D, $W(25)=.918$, $p<.001$; and, E,D&B, $W(28)=.952$, $p=.205$. Therefore, H2 analyses were conducted using seven non-parametric (two-tailed Wilcoxon Signed-Rank Tests) and one parametric (two-tailed Student's t-test) tests. Finally, exploratory analyses are conducted to compare results with previous studies.

Descriptive Findings: Qualitative Judgements. At all four stages the correct qualitative judgement, "It is now more likely the cash was stolen" ('more'), was selected by approximately half of participants (46.7%-56.7%). Conversely, approximately half of participants were also

Table 2. Number and percent of qualitative judgements at each stage of corroborative evidence.

Scenario 1:	**C + A**		**C & A + B**	
	N	%	*N*	%
Correct / 'More'	*28*	46.7%	*32*	53.3%
Incorrect	*32*	53.3%	*28*	46.7%
'Less'	*5*	8.3%	*6*	10.0%
'Same'	*27*	45.0%	*22*	36.7%
Scenario 2:	**E + D**		**E & D + B**	
	N	%	*N*	%
Correct / 'More'	*34*	56.7%	*31*	51.7%
Incorrect	*26*	43.3%	*29*	48.3%
'Less'	*2*	3.3%	*4*	6.7%
'Same'	*24*	40.0%	*25*	41.7%

selecting an incorrect qualitative judgment, both "The probability the cash was stolen remains the same" ('same') and "It is now less likely the cash was stolen" ('less'). A large minority (36.7%-45%) at all four stages selected "The probability the cash was stolen remains the same", only a minority at each stage indicated that their belief should decrease (see Table 2). 27 (45%) participants were consistent in their selections across all four stages; but only a third (n=19, 31.37%) correctly believed that in each instance of corroborating evidence, the likelihood of the target hypothesis should increase. Therefore, approximately two thirds of participants made at least one error when qualitatively evaluating the impact of corroborating evidence. C&A was the only instance in which the incorrect qualitative judgment was more likely to be selected and that 'more' and 'same' were selected at equal rates.

Preferences in Qualitative Judgements (H1). Analysis was conducted to determine if there was a significant preference for the correct qualitative judgement. Number

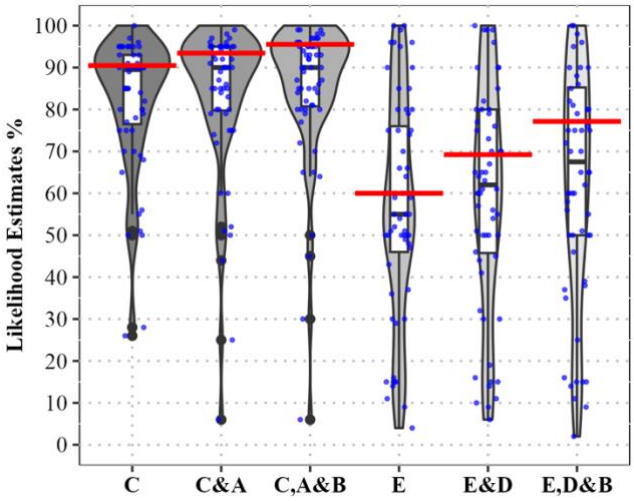

Figure 1. Violin plots showing obtained estimates. Individual estimates (N=60) shown in blue dots, Bayesian predictions shown in red.

and percentage of qualitative judgement selections are shown in Table 2. Four Binomial tests were conducted to test if the proportion of those selecting the correct qualitative judgement ('more') was significantly different from the proportion of those selecting incorrect qualitative judgements, (both 'same' and 'less'). Only one of the tests was found to be significant (p=.050); the correct qualitative judgement was significantly more likely to be selected for E&D. In all other instances there was no significant difference (C&A, p=.184, C,A&B, p=.877 and E,D&B, p=.608), meaning in the majority of cases there was no significant preference for the correct qualitative judgement.

Comparison to Bayesian Predictions by Qualitative Judgement (H2). To determine if group level errors can be explained by qualitative intuitions (as in Phillips et al.,

Table 3: Outcome of H2 analyses. Included are test scenarios, test values, witness combinations medians, outcome of seven Wilcoxon Signed-Rank Tests and one Student's t-test (*). Also shown are 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	Observed value	Test value	Comparison Tests		Effect Size	95% CI Effect Size		VS-MPR
			W or *t*	*p*		Lower	Upper	
Correct								
C&A	92.00	93.44	130	=.098	-.360	-.664	.048	1.612
C,A&B	92.50	95.53	112	=.005	-.576	-.783	-.253	14.871
E&D	65.00	69.23	305	=.905	.025	-.345	.389	1
E,D&B	75.00	77.14	227	=.688	-.085	-.453	.308	1
Incorrect								
C&A	87.50	93.44	63	<.001	-.761	-.885	-.539	246.966
C,A&B	87.50	95.53	17	<.001	-.916	-.963	-.815	1480.261
E&D	50.50	69.23	56	=.003	-.681	-.854	-.372	24.510
E,D&B*	53.52	77.14	-4.423	<.001	-.821	-1.238	-.394	308.115

2023), analysis was conducted to assess if qualitative judgements significantly impact the accuracy of belief estimates. Participant estimates, split by correct or incorrect qualitative judgement, were tested against Bayesian predictions (posterior probabilities, shown in Table 1). The results of these analyses are shown in Table 3. When participants had an incorrect qualitative judgement, all estimates were found to significantly deviate from the Bayesian prediction. However, for those who had a correct qualitative judgment, the only witness scenario found to deviate from the Bayesian prediction was C,A&B, all other estimates approximated Bayesian predictions. In this instance, even those with the correct qualitative intuition that beliefs should increase, failed to adequately adjust their beliefs upwards (median beliefs only increased by 0.5). The results of this analysis indicate that errors in qualitative judgements at the individual level can explain deviations from Bayesian predictions at group level.

Exploratory Analysis. To compare findings with the previous study (Phillips et al., 2023) we will replicate the same group level tests. Obtained belief estimates for each witness combination are shown in Figure 1. In the first scenario (starting with Chris' highly reliable report) estimates are highest for C&A and C,A&B equally (median = 90); marginally higher than estimates for C (median = 89.5). By contrast in the second scenario (starting with Edward's less reliable report), estimates increase at each stage (E median = 55; E&D median = 62; E,D&B median = 67.5). We will examine if at group level, as in the previous study, estimates significantly deviate from Bayesian predictions, with corroborating reports being undervalued. We will also assess whether results indicate that in certain contexts corroborating reports are devalued. These analyses were conducted using non-parametric tests, via JASP. Shapiro Wilk tests found that estimates across all witness combinations violated the assumption of normality: C, $W(59)=.803$, $p<.001$; C&A, $W(59)=.742$, $p<.001$; C,A&B,$W(59)=.733$, $p<.001$; E, $W(59)=.957$, $p=.032$; E&D, $W(59)=.946$, $p=.010$; and, E,D&B, $W(59)=.932$, $p=.002$. The results of these analyses are shown in Table 4.

First, obtained belief estimates are compared to Bayesian predictions across all six witness combinations, with the hypothesis being that participant estimates will be significantly inaccurate. Six two-tailed Wilcoxon Signed-Rank Tests were conducted. It was found that estimates for three corroborative scenarios and estimates for Chris's singular report were significantly inaccurate. Only for E and E&D were inaccuracies found to be non-significant. All obtained estimates were conservative, meaning all witness combinations were undervalued.

Second, belief estimates following a single report and two reports in corroboration are compared. Two one-tailed Wilcoxon Signed-Rank Tests were conducted to determine if corroborating reports are significantly less than (devalued) compared to initial reports. It was found that estimates for C&A were not significantly less than estimates for C, and estimates for E&D were not significantly less than estimates for E. This is however unsurprising given the findings from the qualitative judgments.

Third, belief estimates following two and three reports in corroboration are compared. Two one-tailed Wilcoxon Signed-Rank Tests were conducted to determine if additional corroborating reports are further devalued. Estimates for C,A&B were not significantly less than C&A and estimates for E,D&B were not significantly less than E&D. This again is unsurprising given the findings from the qualitative judgments and the previous analysis.

Table 4: Outcome of exploratory 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). Significant findings, using Bonferroni correction (0.05/10 = 0.005) are indicated by *.

Scenarios	Median	Test value	Wilcoxon Signed-Rank Tests		Effect Size	95% CI Effect Size		VS-MPR
			W	*p*		Lower	Upper	
Comparison to Bayesian Predictions								
C	89.50	90.48	437	<.001*	-.522	-.702	-.281	110.389
C&A	90.00	93.44	382	<.001*	-.583	-.743	-.359	454.628
C,A&B	90.00	95.53	230	<.001*	-.749	-.851	-.591	54404.729
E	55.00	60.00	703	=.441	-.119	-.398	.180	1
E&D	62.00	69.23	640	=.043	-.301	-.538	-.019	2.707
E,D&B	67.50	77.14	512	=.003*	-.440	-.643	-.180	20.901
Two corroborating reports are not devalued								
C&A < C	90.00	89.50	743	=.103	-.188	-	.054	1.572
E&D < E	62.00	55.00	1105.5	=.952	.249	-	.463	1
Further corroborating reports are not devalued								
C,A&B < C&A	90.00	90.00	624.5	=.079	-.217	-	.032	1.837
E,D&B < E&D	67.50	62.00	1003	=.815	.133	-	.363	1

Discussion

This study sought to further examine the reasoning errors identified in Phillips et al. (2018; 2023); to investigate the accuracy of qualitative intuitions and belief estimates when evaluating the added value of supplementary evidence in the form of corroborative reports.

Qualitative judgements in this study were directly measured rather than inferred, therefore we were able to explore propensity to make qualitative errors and if these errors were significant. We find that two thirds of participants made at least one qualitative error when evaluating the impact of corroborating evidence. In three of four instances participants were equally likely to select correct or incorrect qualitative judgements, meaning there was no significant preference. The only instance in which there was a significant preference for the correct qualitative judgement (that beliefs should increase) was for E&D; qualitative errors were significantly less likely when a credible report of low reliability is corroborated by a report of equal reliability. Conversely, qualitative errors were most likely to be made when a highly reliable report is supplemented by a less reliable but credible report (C&A), this was the only instance in which more participants selected an incorrect than correct response (53.3% vs. 46.7%) and participants were almost equally likely to indicate that beliefs should increase or stay the same (46.7% vs. 45.0%). This was the simplest test of this experimental paradigm so far, compared to previous studies where participants were asked to rank or provide belief estimates, participants were simply asked if additional credible evidence should impact beliefs (increase, decrease or remain the same). Furthermore, evidence was sequentially presented, and participants were explicitly prompted at each stage to evaluate impact. It is surprising that even still, errors are made at such a high rate, revealing robust erroneous qualitative intuitions.

Following qualitative judgements belief estimates were also measured, to be able to explore whether sub-groups with differing qualitative intuitions can explain the inaccuracies in belief estimates observed at group level, as proposed based on exploratory findings in Phillips et al. (2023). Expectedly, we confirmed that those who selected incorrect qualitative judgments gave significantly inaccurate belief estimates. Likewise, we were also able to confirm that those who selected correct qualitative judgements in most instances provided accurate belief estimates, approximating Bayesian predictions. There was only one instance (C,A&B) in which those with the correct qualitative intuitions failed to adjust their beliefs accordingly; despite indicating that their beliefs should increase beliefs were adjusted minimally upward. Therefore, we also conducted exploratory analysis to assess whether group level findings can be replicated from previous studies (Phillips et al., 2023). All obtained estimates were conservative, meaning, the value of eyewitness reports were consistently underestimated. Belief estimates for three corroborative scenarios and for Chris's singular report significantly deviated from Bayesian predictions. Based on the above findings, error in belief estimates seen at the group level can be explained by differences in qualitative intuitions at the individual level. Unlike the previous study we did not find that corroborating reports were devalued, participants did not significantly decrease their estimates when introduced with further corroborating evidence. This is however unsurprising given the findings from the qualitative judgments; as only a handful of participants (3.3-10%) indicated that their beliefs should decrease, whereas in the previous study participants were much more likely to decrease their beliefs (31.7-41.7%). In this study, those who made incorrect qualitative judgments were much more likely to indicate that their beliefs should stay the same (36.7-45.0%), whereas in the previous study only a minority of participants kept their beliefs the same (16.7-25%).

Findings demonstrate that large proportions of individuals have a general inability to accurately evaluate the value of corroborating evidence, however, findings also indicate a specific intuitive error that causes individuals to inaccurately incorporate additional corroborating evidence from sources of lower reliability, i.e. a selective difficulty in accurately integrating both cues of reliability and number of reports. These identified errors indicate the operation of alternative reasoning strategies. Those who indicate that beliefs should decrease may be using intuitive strategies such as averaging (e.g., Lopes, 1985) or dilution (Madsen, Hahn, & Vorms, 2017). Those who indicate that beliefs should remain the same may be using intuitive strategies that correspond to the predictions of the MAXMIN rule (see Walton, 1992, 2007). However, a complete explanation cannot yet be offered for this error, further work is needed to explore possible alternative reasoning strategies.

Conclusions

Overall, this study concurs with previous findings that lay reasoners do not integrate corroborative testimonies in the manner expected by normative, Bayesian, predictions. Large sub-populations show evidence of flawed qualitative intuitions when evaluating the impact or combined value of corroborating evidence; as shown by inaccurate qualitative judgements and belief estimates. Erroneous intuitions are particularly evident when it is necessary to assess both informational cues of reliability and report number; intuitions are led further astray when initial reports are of high reliability and additional credible corroborative reports are from sources of lower reliability. This surprising finding demonstrates that even in a relatively simple paradigm, with a simple qualitative experimental task, and when explicitly prompted at each stage, lay reasoner intuitions still result in significant errors. However, further work is needed to determine explanations for these identified errors.

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. https://doi.org/10.1037/h0076477

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.

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 Academic

Durkin, K., & Shafto, P. (2016). Epistemic trust and education: effects of informant reliability on student learning of decimal concepts. *Child development*, *87*, 154-164.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart*. 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*. 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].

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.

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

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. Bulletin of the Psychonomic Society, 23, 509–512.

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 (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* Conference of the Cognitive Science Society.

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.

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).

Phillips, K., Hahn, U., & Pilditch, T. D. (2023). Evaluating testimony from multiple witnesses: consistent undervaluing and selective devaluing of corroborating reports. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 3338-3344).

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., & 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.