Getting it Right: Communication, Voting, and Collective Truth Finding

Valeria Burdea[*] LMU Munich

Jonathan Woon[†] University of Pittsburgh

January 18, 2023[‡]

Abstract

We conduct an experiment in which groups are tasked with evaluating the truth of a set of politically relevant facts and statements, and we investigate whether communication improves information aggregation and the accuracy of group decisions. Our findings suggest that the effect of communication depends on the underlying accuracy of individual judgments. Communication improves accuracy when individuals tend to be incorrect, but diminishes it when individuals are likely to be correct ex ante. We also find that when groups vote independently without communicating, subjects update their beliefs in a manner consistent with interpreting others' votes as mildly informative signals, but not when they communicate beforehand. The chat analysis suggests that group members use communication to present their knowledge of related facts and to engage in interactive reasoning. Moreover, the volume of both types of communication increases with item difficulty.

Keywords: collective decisions, voting, communication
JEL Classification: D70, D72, D83

[*]valeria.burdea@econ.lmu.edu. Assistant Professor, Department of Economics, LMU Munich

[†]woon@pitt.edu. Professor, Department of Political Science, Department of Economics (secondary), and Pittsburgh Experimental Economics Laboratory

[‡]Previous versions were presented at the 2019 Annual Meeting of the American Political Science Association, Washington, D.C and at Florida State University. The University of Pittsburgh Central Research Development Fund provided funds for this research, and the University of Pittsburgh IRB approved the study as an exempt protocol (STUDY19030295)

1 Introduction

To make good decisions, groups, organizations, and societies often need to correctly assess the facts. A hiring committee may be tasked with determining whether a job candidate has the right qualifications and experience. A jury must decide whether the evidence they hear is reliable in rendering a verdict on a defendant's guilt or innocence. An electorate considers an incumbent's record to gauge whether they are corrupt or public spirited. A community considers the environmental impact against the economic benefits of a new business development. Getting the facts right in any of these situations may be a difficult problem for any single individual, because of incomplete information or because the evidence leads to the formation of only a vague opinion. But when groups make decisions, their collective judgments have the potential to be much more accurate than individual judgments.

One mechanism through which group judgments can be superior is by statistical aggregation, whereby combining diverse viewpoints or pieces of information results in the "wisdom of the crowds" (Galton, 1907; Kelley, 1925; Stroop, 1932; Surowiecki, 2004). Majority voting can also be an effective aggregation mechanism, an idea expressed in the Condorcet Jury Theorem. As long as individual opinions are generally correct, not only does the majority opinion have a greater chance of being correct than any individual, the majority opinion will almost certainly be correct for groups that are large enough (Austen-Smith and Banks, 1996).[1] This "truth-producing property" of majoritarian decisions provides the basis for those who justify and value democracy for its epistemic virtues (Landemore, 2017; List and Goodin, 2001).

Research on team reasoning and collective intelligence (Cooper and Kagel, 2005; Kugler, Kausel and Kocher, 2012; Navajas et al., 2018; Woolley et al., 2010) also suggests the superior accuracy of group decisions. Indeed, team reasoning necessarily involves communication between team members. Similarly, members of real-world juries do not simply vote,

[1] If individual opinions tend to be incorrect or biased, then majority voting has a "dark side" whereby groups can end up being worse than individuals (Morton, Piovesan and Tyran, 2019).

they deliberate, and communication is necessary for deliberation, involving the exchange and evaluation of facts and arguments (Mercier and Sperber, 2017). But communication does not guarantee superior collective judgment. Groups might engage in groupthink if members prefer conformity and cohesion to accuracy (Janis, 1982). They might also be dominated by their least informed, but most overconfident, members. Or the groups themselves might become overconfident and less accurate if they fail to discount commonly held information, much like in an echo chamber (DeMarzo, Vayanos and Zwiebel, 2003; Lorenz et al., 2011; Lorenz, Rauhut and Kittel, 2015; Jasny, Waggle and Fisher, 2015). Whereas the effectiveness of majoritarian information aggregation relies on the independence of opinions, deliberation trades statistical independence for enhanced reasoning.

Do groups form more accurate judgments when they can communicate before voting than by voting alone? We designed and conducted an incentivized experiment to investigate the effect of communication on the accuracy of small-group majority opinions regarding a variety of real-world, politically-relevant facts.[2] Our research is therefore contributing to the literature on collective decision-making. The most related studies were conducted by Lorenz, Rauhut and Kittel (2015), Morton, Piovesan and Tyran (2019) and Goeree and Yariv (2011). Lorenz, Rauhut and Kittel (2015) vary the decision rule (individual, majority, unanimity), group size, and the mode of communication (qualitative or quantitative), but did not investigate the effect of communication per se (holding constant the possibility of communication). They found that group judgments were worse with majority rule. This is at odds with the findings of Goeree and Yariv (2011) where a simple majority rule is at least as efficient as a unanimity rule both with and without communication. One main difference between these studies is the task: in the former study, participants were confronted with general knowledge questions with a larger (0-100) answer space, while in the latter,

[2]This context is important because an informed citizenry is widely believed to be a critical component of a healthy democracy (Lupia and McCubbins, 1998). However, extensive research has documented that citizens are not only poorly informed, but quite often misinformed (Berinsky, 2015; Delli Carpini and Keeter, 1996; Gilens, 2001; Kuklinski et al., 2000). Accuracy and learning have been shown to be hampered by confirmation bias and motivated reasoning, often the result of intense partisanship (Bartels, 2002; Gaines et al., 2007; Jerit and Barabas, 2012).

participants had to predict the outcome of a binary variable in an abstract, balls and urns, task. Given the common interest in the effect of communication of collective decisions, our study is closer to Goeree and Yariv (2011). However, one of the important differences between our study and theirs is the scope of communication due to the task. Our context laden environment gives us more room to differentiate between two channels through which communication can impact collective decisions: the interactive reasoning and information exchange channels which may lead to opposite effects compared to a no-communication protocol.

The experiment conducted by Morton, Piovesan and Tyran (2019) did not allow for communication but was instead aimed at testing whether providing social information could "debias" voters. They find that such information had no effect. Mercier and Claidière (2022) provide an overview of several studies suggesting that discussion in small groups improves these groups' performance, and they are interested in the effect of communication on decisions in large groups (N greater than 22) using a within-subject design where learning effects may play an important role. Although these studies vary in the type of problems groups are faced with, they do not differentiate between difficulty levels across problems. Our study complements the literature in this dimension and differs in that our focus is on small groups, where dynamics may be significantly different.

In addition to observing individual votes and group decisions, as in previous studies, our design is innovative in that we also measured individuals' prior and posterior beliefs using an incentivized belief elicitation procedure. This feature of our design allows us to assess how individuals' beliefs changed as a consequence of communication and group decisions. We find that communication does improve the accuracy of group decisions, but the effect is not uniform. Communication improves group accuracy relative to voting independently for hard cases in which individuals tend to be incorrect, but diminishes it for easy cases in which individuals are already likely to be correct. Using our data, we estimate an empirical belief-updating model and find that beliefs change in response to the prior information brought

by other group members only when communication is not permitted, and we also find that posterior beliefs are only more accurate with communication for hard items. Finally, our analysis of communication transcripts suggests that subjects do not simply announce their vote intentions, but communicate their knowledge of related facts, raise doubts, and attempt to reason together towards the truth.

2 Theoretical Framework

We consider a setting in which a group votes by majority rule to decide whether some claim is true or false. This situation is analogous to a jury determining whether a defendant in a criminal trial is guilty or innocent or a legislative committee deciding whether a particular bill is good or bad policy. Let p_i be the probability that an individual's vote is accurate, and let p_g denote the probability that a group decision (i.e., a majority of votes) is accurate. As long as individuals are more accurate than chance ($p_i > \frac{1}{2}$) and votes are statistically independent, then the logic of the Condorcet Jury Theorem implies that groups are more likely to be accurate than individuals, $p_g > p_i$. This follows from a straightforward application of the binomial probability distribution. However, if individuals tend to be less accurate than chance ($p_i < \frac{1}{2}$), then the flip side of aggregation by majority rule is that groups will be *less* accurate than individuals, $p_g < p_i$. The relationship between group and individual accuracy is depicted by the solid black line in Figure 1, which is above the gray 45 reference line for $p > \frac{1}{2}$ and below it for $p < \frac{1}{2}$. Note that this simple framework ignores why individuals may be more or less accurate, though we presume that different individuals have access to different pieces of information that, on average, would imply the overall level of accuracy p_i.

How might communication prior to voting affect the quality of a group's decision? If the process of communication elicits no new information from group members—for example, they simply announce their votes (with the same probability of accuracy p_i), but no one changes their opinion—then we would expect group decisions reached with and without

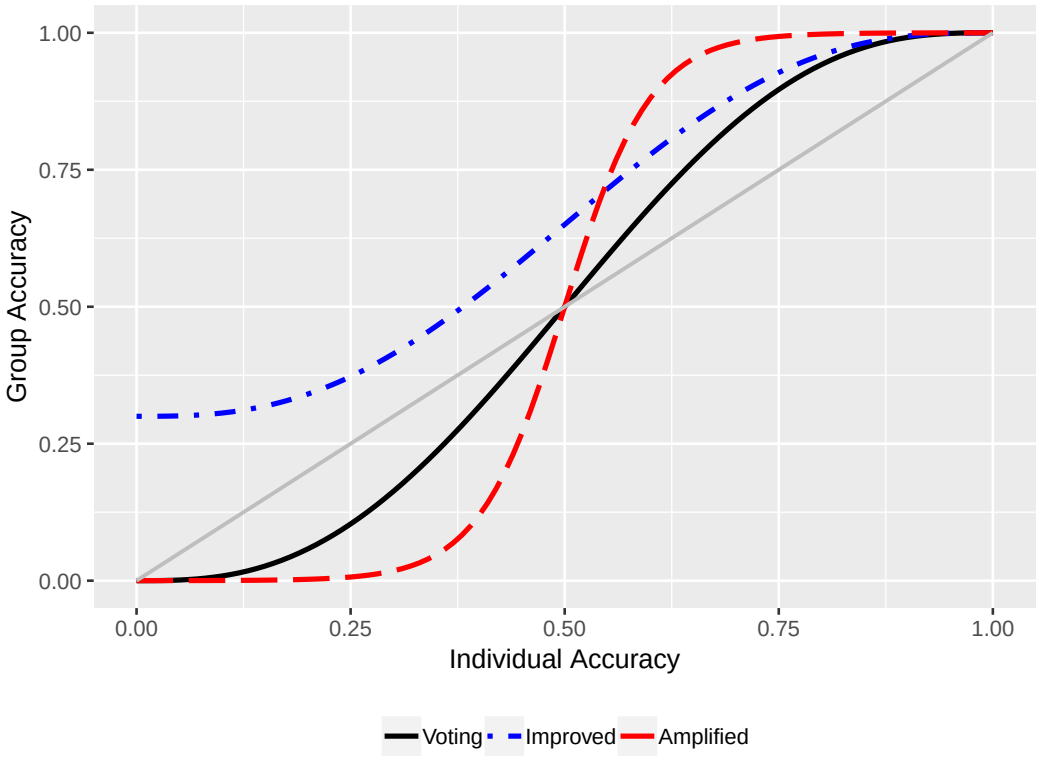


Figure 1: Group versus individual accuracy and the hypothesized effects of communication

communication to have the same degree of accuracy. If individuals announced their vote intentions and then voted with the majority opinion, then communication would affect the degree of group consensus (all votes would be unanimous) without affecting accuracy. In both cases, the relationship between group and individual accuracy would still look like the solid black line in Figure 1.

Communication might increase group accuracy if it facilitates *information exchange*. Consider a silly, but simple, example in which a group is trying to determine whether a particular animal is a duck. One member knows only that the animal swims, while another member knows only that the animal flies. On its own, each piece of information is not terribly informative, but by sharing their information, their knowledge that the animal both swims and flies increases their chance of correctly identifying the animal as a duck (though they could still think the animal is some other water bird).

Group accuracy might also increase if communication facilitates processes of deliber-

ation and *interactive reasoning* (Mercier and Landemore, 2012; Mercier and Sperber, 2017). We can think of deliberation as involving an exchange of arguments in which facts and conclusions are critically examined, challenged, and investigated before they are accepted. Deliberation is one form of team reasoning—a general process in which communication allows groups to pool cognitive resources so they are better able to work out the implications of the information they've collected. Team reasoning could operate independently of information exchange, if individuals have the same information but are more likely to work out the implications together than alone, or it could operate in a way that complements and enhances information exchange. Thus, reasoning together is distinct from information aggregation as a mechanism that increases group accuracy.

We hypothesize that when communication facilitates interactive reasoning (either with or without information exchange), group accuracy will be greater when groups communicate before they vote than when they make decisions by voting alone. This *improvement* hypothesis is illustrated by the blue dashed-dotted line in Figure 1 that lies above the solid black line regardless of whether individuals tend to be accurate or inaccurate. Implicit in this hypothesis is that deliberation and interactive reasoning can enable groups to recognize when their individual information is biased (i.e., inaccurate) and to correct for this bias. We also note that although the improvement in accuracy shown in the figure is merely for illustration purposes (as it is not intended to represent a quantitative prediction), gains in accuracy are necessarily smaller when individuals are more accurate due to a ceiling effect.

If communication enhances information exchange without improving interactive reasoning, we hypothesize that instead of universally improving group accuracy, it may *amplify* the effects of majority rule aggregation. This could happen when each individual brings a piece of information that points in the same direction that, when combined, implies greater confidence in the original conclusion. That is, when individual information tends to be accurate, groups are more likely to make the correct decision after communicating with each other, but are less likely to be accurate if members' individual information is inaccurate to

begin with. This possibility is shown by the dashed red line in Figure 1 that lies above the black line for $p_i > \frac{1}{2}$ and below it for $p_i < \frac{1}{2}$.

3 Experimental Design and Procedures

The central task in the experiment was for groups to determine whether real-world statements were true or false. We selected a total of 32 statements as items to evaluate in the experiment, divided into two sets of 16 statements that we designate sets F and P.[3] In each set, half of the statements were true and the other half were false. Set F statements were *factual*, involving 12 policy-relevant facts similar to the kinds of items used in political knowledge studies (Bullock et al., 2015; Hill, 2017; Robbett and Matthews, 2018), such as changes in the unemployment rate, median household income, number of refugees, or gun-related deaths. Following Hill (2017), most policy-relevant facts in set F were worded in such a way that implied a partisan-favorable direction if the statement were true. For example, the statement "The total public debt of the United States federal government more than doubled from quarter 2 in 1981 to quarter 1 in 1989 while Ronald Reagan was president" is favorable to Democrats, while "Since President Trump took office in 2017, the civilian unemployment rate has decreased by almost 1 percentage point" is favorable to Republicans. For comparison purposes, we also included 4 additional facts that were neither partisan nor policy-relevant (three historical facts and one about airports).

Set P statements involved claims made by real *politicians*, such as Hillary Clinton, Nancy Pelosi, and Marco Rubio. We selected these statements because we are interested in whether citizens are able to evaluate the kinds of claims made in the course of ordinary politics, heard in candidates' and elected officials' speeches or reported in the news. Because the truth of such statements is typically more subjective in nature, we restricted our attention to policy-relevant claims that could be researched and fact-checked. Following Woon (2019),

[3]The full list of statements is provided in the Appendix. The statements were selected based on (unincentivized) pre-test results to ensure sufficient variation in individual accuracy while also balancing their truth and partisan direction.

we relied on PolitiFact to assess whether the statements were true or false.[4]

We selected the two types of statements to construct a more diverse pool of items. In our analysis we always control for individual item characteristics in order to identify more general patterns of heterogeneity in the effect of communication. Nevertheless, we also check for differences between the two groups of statements and discuss these in the Results section.

We first describe the overall design of the experiment before providing a detailed explanation of our procedures. Each session of the experiment was divided into separate parts, and each part consisted of 16 rounds. Each round within a part corresponded to a different statement to be evaluated. In Part 1 of every session we elicited *prior beliefs* about the truthfulness of each statement. Part 2 was the *group decision task* with decisions made by majority rule. In Part 3 we elicited *posterior beliefs*. The order of statements was randomized at the individual level in Parts 1 and 3. In Part 2, since the group decision required all group members to see the same statement at the same time, the order of statements was randomized at the session level.

We varied the possibility of communication prior to voting so that communication was not possible in *Voting* (only) sessions, while group members could communicate with each other via free-form messages prior to casting their votes in each round of the *Chat* sessions. Within every session, Parts 1-3 involved only F statements or P statements, but not both. In Voting sessions, Parts 4-6 repeated the sequence of prior belief elicitation, group decision task, and posterior belief elicitation for the set of statements not rated in Parts 1-3. Thus, in Voting sessions, subjects voted on both F and P statements, with either F statements in Parts 1-3 and then P statements in Parts 4-6 or vice versa. In Chat sessions, due to the additional time required for communication, subjects only rated one type of statement. Overall, there were four types of sessions that varied according to whether communication was allowed and the order or type of statements that were rated. The structure of the

[4]PolitiFact has a six-point scale to rate the truthfulness of statements, and we only selected statements that they deemed to be unambiguously true or false, avoiding ambiguous "mostly true," "half true," and "mostly false" statements. We did not include any statements by President Trump due to their distinctive and recognizable linguistic properties.

Table 1: Summary of experiment

			Session Type			
Part	Task	Rounds	VF	VP	CF	CP
1	Prior beliefs	16	Beliefs F	Beliefs P	Beliefs F	Beliefs P
2	Group decision	16	Voting F	Voting P	Chat F	Chat P
3	Posterior beliefs	16	Beliefs F	Beliefs P	Beliefs F	Beliefs P
4	Prior beliefs	16	Beliefs P	Beliefs F	–	–
5	Group decision	16	Voting P	Voting F	–	–
6	Posterior beliefs	16	Beliefs P	Beliefs F	–	–
	Number of sessions		3	3	3	4
	Number of subjects		60	55	60	75

experiment is summarized in Table 1. Subjects accumulated earnings in each part, and we randomly selected one part to count for payment once all of the decision tasks in the session were completed. Before being paid, subjects also completed a brief questionnaire.

The group decision task was straightforward and generated the primary outcome of interest, which was whether or not groups made accurate decisions. In each round, subjects were randomly assigned to groups of five, with random reshuffling of groups prior to every round. Subjects voted on whether the statement was true or false, without the possibility of abstention. The group decision was determined by majority rule and every group member earned a bonus of \$1 if the group's decision was correct and \$0 otherwise.[5] To prevent possible learning effects (e.g., about others' or one's own accuracy), no feedback was given about the number votes, earnings, or the accuracy of decisions until after the conclusion of all rounds of the group decision task. While we provided information about the number of votes and the group decision when we elicited posterior beliefs, no feedback about accuracy or earnings was provided until the conclusion of the experiment.

[5]Subjects were instructed that when they rated statements from PolitiFact they earned the bonus if the group's decision matched PolitiFact's evaluation.

In Chat sessions, group members had 3 minutes to communicate with each other using free-form text messages in a standard chat room format. Group members had one chat window and could see the complete history of messages within their group during each specific round but could not communicate with members of any other group. We did not give any instructions for how subjects should use the chat other than to ask them to refrain from providing identifying information about themselves and to avoid using offensive language.[6]

Our secondary outcome of interest was the change in beliefs before and after group members were exposed to social information. To elicit probabilistic beliefs about the truth of each statement, we used the crossover elicitation method in the parts before and after the group task.[7] We chose the crossover method both for its desirable theoretical properties (it is incentive compatible regardless of risk preferences) and because, among the many belief elicitation methods proposed in the literature, it is possible to explain the mechanics of the task and to demonstrate its incentive compatibility to subjects without the need to invoke mathematical formulas, instead using only simple ordinal comparisons.

The crossover elicitation procedure in our experiment worked as follows. In each round, subjects were asked to choose a number B from 0 to 100 that represented their "belief in the likelihood the statement is TRUE" by dragging a slider on the screen. In the instructions, we explained the meaning of different values of B: that 100 represented complete certainty the statement is true, 0 represented complete certainty the statement is false, and 50 represented total uncertainty.[8] Subjects could earn a bonus of \$1 depending on the value of a random number W in relation to their belief B and whether or not the statement was indeed true. Specifically, we explained that W represented the number of "winning lottery tickets" in the round and that its value was drawn uniformly from 1 to

[6]We also gave groups the option of exiting the chat room before 3 minutes expired if its members unanimously agreed. We included this feature so that subjects would not feel compelled to communicate and could make their decision more quickly if they had no information to share.

[7]The crossover method is a stochastic version of the BDM mechanism and was first used by Allen (1987) and Grether (1992), and subsequently analyzed theoretically by Karni (2009). It has since been used in experimental economics and political science by Mobius et al. (2011), Holt and Smith (2016), Hill (2017), and Woon and Kanthak (2019).

[8]After providing a complete explanation, we also provided subjects with a one page summary of the elicitation task (included in the Appendix).

100. If $B \geq W$, then they earn the bonus if the statement is true, while if $B < W$, then they earn the bonus with probability $\frac{W}{100}$. The belief B is therefore the "crossover" point at which a subject (with standard von Neumann-Morgenstern preferences) will be indifferent between the two lotteries (the "subjective" lottery associated with their belief B or the lottery associated with the objective probability W). The method is incentive compatible because if beliefs are misreported (in either direction), there will always exist values of W such that a subject would prefer to switch from one lottery to the other (e.g., from the subjective to the objective or vice versa). Reporting B "honestly" ensures that a subject will be paid according to the lottery that has the higher probability of receiving the bonus (except for $B = W$, when the subject is indifferent).[9]

We refer to the belief elicited in the part before the group task as the prior belief because it is the initial belief before any group interaction. We set the default position of the slider for each statement to $B = 50$ for the prior belief elicitation tasks in Part 1 (and Part 4 in the Voting sessions). When we elicit beliefs in the part following the group decision task, we provide information about how group members voted, but not the accuracy of the group's decision. Thus, the second elicitation is the posterior belief after being exposed to social information: how others' voted (in both Vote and Chat sessions) and how groups communicated (only in Chat sessions). In Part 3 (and Part 6), we set the default position of the slider for each statement to the subject's own prior belief (elicited in Parts 1 or 4, respectively).

[9]We note that discussions of belief elicitation procedures in the experimental literature tend to be concerned primarily with encouraging the "honest" reporting of beliefs and therefore with the problem of incentive compatibility. However, this assumes subjects *have* beliefs and that such beliefs are meaningful for decisions in the ways consistent with mathematical models of decision making under risk and uncertainty. In this view, subjects could report their belief accurately if they wanted to, but either do not want to be honest or do not want to spend the effort be accurate. For example, if a subject's "true" belief is 0.62, they might report 0.5, 0.6, or 0.7 when asked on a survey. A different problem, however, may be that subjects have beliefs but do not know how to express them in quantitative terms. The crossover method is then useful not only because it is incentivized, but also because it provides a means by which subjects can quantify their beliefs by relating them to monetary gambles. In other words, incentive compatability is normally thought of as solving the problem of dishonest reporting, but it can also be thought of as solving the problem where subjects do not know how to express B until it is equated with gambles (assuming they have well-defined preferences over gambles).

The experiment was conducted at the Pittsburgh Experimental Economics Laboratory between April 12-23, 2019. University of Pittsburgh undergraduates were recruited from the lab's general subject pool, and 250 subjects participated in 13 sessions of the experiment (15-20 subjects per session). All interactions between subjects took place via networked computers using an interface programmed in z-tree (Fischbacher, 2007). Participants were on average 20 years old and 65% of them were female.[10] Each session lasted approximately 90 minutes and average payments amounted to $16.

4 Results

4.1 Accuracy

We begin our analysis by examining the accuracy of individual and group decisions. Aggregating across all statements in our data, we find that individuals vote correctly (for true when the statement is true, and for false when the statement is false) 52.3% of the time in the Vote treatment and 59.7% of the time in the Chat treatment (the difference is statistically significant, $p < 0.001$, two-tailed). The accuracy of decisions at the group level across the two treatments mirrors that at the individual level, with 53.1% correct decisions in Vote and 59.7% correct decisions in Chat ($p = 0.027$). This broad, bird's eye view of the data therefore suggests that communication increases accuracy.

When we disaggregate the data, however, a more nuanced picture emerges in which differences in accuracy across the treatments depend on question difficulty. To measure question difficulty, we need a measure of individual accuracy that is not affected by communication, so we construct a measure of *belief-based accuracy* using the prior beliefs we elicited in parts 1 and 4 of the experiment. Specifically, for the prior belief that the statement is true, denoted $p_{ij} \in [0, 1]$ for subject i and item j, belief-based accuracy a_{ij} is an indicator that is 1 if $p_{ij} > \frac{1}{2}$ for true statements or $p_{ij} < \frac{1}{2}$ for false statements and 0 otherwise. That

[10] For a full list of sample demographics, see the Appendix.

Table 2: Individual and group accuracy by treatment and item difficulty

Accuracy Measure	Easy		Medium		Hard		Total	
	Vote	Chat	Vote	Chat	Vote	Chat	Vote	Chat
Prior beliefs	0.67	0.68	0.52	0.52	0.39	0.39	0.52	0.52
Posterior beliefs	0.73	0.73	0.54	0.55	0.36	0.43	0.53	0.57
Individual votes	0.77	0.81	0.52	0.57	0.33	0.44	0.52	0.60
Group decisions	0.92	0.83	0.52	0.56	0.22	0.44	0.53	0.60
N beliefs/votes	1,035	630	1,380	825	1,265	705	3,680	2,160
N groups	207	126	276	165	253	141	736	432

is, we count a subject's belief as being accurate as long as it is in the correct direction, treating complete uncertainty as an incorrect belief. We then computed the average belief-based accuracy at the item level and divided our knowledge items into three categories: *easy* items (average accuracy over 0.6), *medium* items (accuracy between 0.4 and 0.6), and *hard* items (accuracy less than 0.4).

Table 2 presents several measures of accuracy disaggregated by question difficulty and treatment. In addition to the accuracy of individual votes and group decisions, we also examine the accuracy of both prior and posterior beliefs.[11] Looking at prior beliefs serves as a randomization check, as we would not expect any differences between Vote and Chat when items are aggregated by difficulty, and this is indeed what we find. Interestingly, when we look at the accuracy of posterior beliefs, we see that there is an overall increase in accuracy in Chat compared to Vote (0.53 in Vote and 0.57 in Chat, $p < 0.001$), which mirrors the increase in the accuracy of individual votes and group decisions due to Chat. When posterior beliefs are disaggregated by question difficulty, however, the data suggest that the increase in accuracy comes almost entirely from hard questions, with average posterior beliefs for easy and medium items virtually unchanged.

[11] Our continuous measure of the accuracy of a belief is p_{ij} for true statements and $1 - p_{ij}$ for false statements.

Next, we note there are intriguing differences in how changes in accuracy depend on item difficulty for beliefs, votes, and decisions. Although posterior beliefs are more accurate only for hard items, we find that individual votes are more accurate in Chat than Vote across *all* levels of question difficulty. For group decisions, however, the story is a bit different. We find a fairly sizable increase in accuracy due to communication for hard questions (from 0.22 to 0.44), a modest increase for medium questions (from 0.52 to 0.56), and a *decrease* in accuracy for easy questions (from 0.92 to 0.83).

We analyze these differences more systematically by estimating a model for each accuracy measure as a function of question difficulty (indicators for easy and hard items, with medium items as the baseline), the chat treatment, and their interactions. The models also include a set of random effects to account for additional unobserved heterogeneity at the item and session level (for individual-level and group-level dependent variables) and for individual subjects (for individual-level variables). The results of this analysis are presented in Table 3 and provide rigorous support for the basic conclusions drawn from the disaggregated data in Table 2. There are no significant differences in prior beliefs for any level of item difficulty (column 1), and posterior beliefs are significantly more accurate in Chat than Vote only for hard items (column 2). In terms of the accuracy of individual votes (column 3), the main effect of chat is significant and the estimated increase in accuracy is 0.058. While the interaction is positive for easy items and negative for hard items, neither is statistically significant, supporting our conclusion that allowing communication increases individual accuracy irrespective of item difficulty. For group decisions (column 4), there are clear conditional effects of communication, as the main effect of Chat is positive but not significant, while the coefficients for the interactions are fairly large in magnitude, significant, and in opposite directions (-0.140 for easy items and 0.175 for hard items). Our analysis clearly suggests that although individual votes become more accurate with communication across all levels of item difficulty, this does not translate into uniformly better group decisions. Communication helps groups perform better only for medium and hard questions, and leads to worse

Table 3: Effect of chat depends on item difficulty

	Dependent variable:			
	Prior Belief	Posterior Belief	Individual Vote	Group Decision
	(1)	(2)	(3)	(4)
Chat	0.001	0.017	0.058	0.040
	(0.012)	(0.019)	(0.028)	(0.042)
Chat × Easy	0.006	−0.018	−0.016	−0.140
	(0.018)	(0.019)	(0.031)	(0.063)
Chat × Hard	0.002	0.056	0.041	0.175
	(0.017)	(0.019)	(0.030)	(0.060)
Easy	0.152	0.194	0.253	0.401
	(0.025)	(0.039)	(0.049)	(0.067)
Hard	−0.125	−0.181	−0.187	−0.300
	(0.024)	(0.037)	(0.046)	(0.064)
Constant	0.516	0.537	0.516	0.522
	(0.017)	(0.028)	(0.035)	(0.044)
Observations	5,840	5,840	5,840	1,168
Log Likelihood	−612.026	−961.115	−3,769.208	−675.363

*p<0.1; **p<0.05; ***p<0.01

Linear mixed effects model with subject, item, and session random effects in columns 1-3, and item and session random effects in column 4.

group performance for easy ones. This effect is robust to excluding extremely hard questions (average belief-based accuracy less than 0.2) and extremely easy ones (average belief-based accuracy greater than 0.8) suggesting it is not a simple boundary effect (see Table C8 in Appendix for the corresponding regression analysis). The effect is also directionally unaffected by separating between the type of statement (P or F), except that when we do so, the positive effect of chat for Hard items is only significant for P statements, and the negative one for Easy items only significant for F statements (see Tables C3 and C4 in the Appendix). This is due to an imbalance in the distribution of Easy and Hard items across these groups of statements: for P statements the amount of Easy items are approximately double that of Hard items while for F statements the opposite is the case (see Tables C1 and C2 in the Appendix).

Result 1. *Communication improves individual accuracy for all levels of item difficulty but its effect on group performance depends on item difficulty. Communication improves group accuracy for items of medium and hard difficulty, and worsens it for easy items.*

We generate the empirical analogue of Figure 1 by disaggregating the data further, to the item level, and plotting group accuracy against individual (belief-based) accuracy. The results are shown in Figure 2. The horizontal axis is an item's average belief-based accuracy, which we use as a measure of underlying individual accuracy (the likelihood that any individual correctly identifies whether the statement is true or false), and the vertical axis shows the proportion of correct group decisions. In addition to the item-level observations, we also fit loess curves for each treatment and a reference curve (in black) showing the probability of a correct majority decision calculated as a function of the binomial distribution. The relationship between individual and group accuracy in the Vote treatment (blue triangles and loess curve) follows an S-shaped pattern that is consistent with the statistical aggregation of independent votes, which is what we would expect in the absence of communication.[12]

[12]Interestingly, the blue loess curve appears to be slightly above the black reference curve. This may be an artifact of our coding uncertain beliefs as incorrect, because by inducing a conservative (downward) bias in belief-based accuracy, the overall curve shifts to the left.

Although the groups in our experiment are small, this pattern is also consistent with groups being more accurate than individuals when individuals have a better than chance probability of being accurate and groups being worse than individuals when individuals are worse than chance.

We also see in Figure 2 a more detailed version of the basic pattern for Chat (red circles) we noticed from the aggregate statistics in Table 2 and the models in Table 3. The loess curve for Chat (red curve) is above the loess curve for Vote on the left side of the figure and below it on right side of the figure. That is, group accuracy improves with communication when individuals are unlikely to be correct (hard items) and declines when individuals are already likely to be correct (easy items). Although the observed effect of communication is consistent with our expectations for hard statements (as in both of the hypothesized effects in the left side of Figure 1), we did not expect communication to decrease performance.

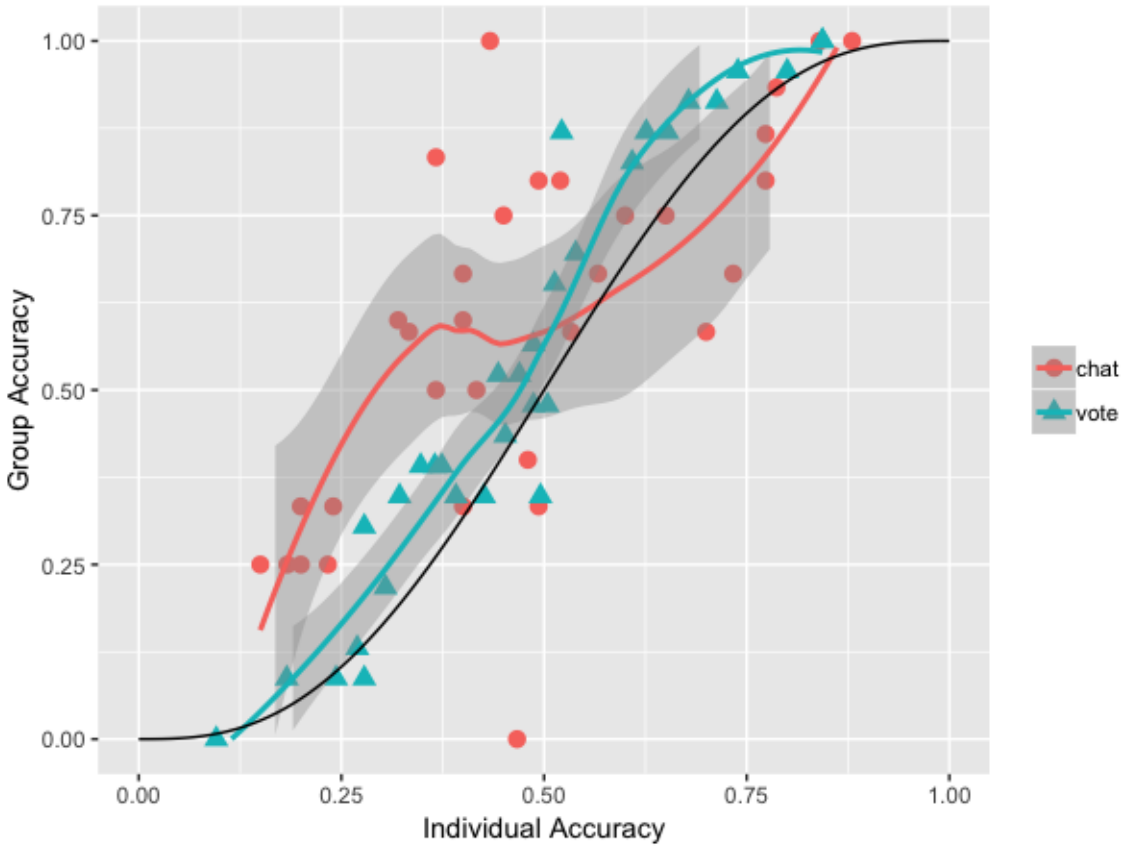


Figure 2: Group versus individual accuracy

One possible explanation for this unexpected finding is that communication affects group performance because it increases skepticism or uncertainty. That is, instead of exchanging different pieces of information that the group can aggregate or scrutinize to arrive at a more reasoned and accurate judgment than is possible through the mere exchange of simple opinions, communication allows group members to express doubt to one another. By increasing doubt, groups will then be less certain their decisions are correct, thereby decreasing confidence and leading groups to make decisions closer to random guesses (that is, decreasing the slope of the relationship between group and belief-based accuracy). For easy items, guessing decreases accuracy, but increases it for hard ones.

While this pattern is consistent with the differences between Vote and Chat at the extremes, it does not quite fit what we observe in the middle range. For many medium-difficulty items on the inaccurate side of the horizontal axis (those with individual accuracy between approximately 0.3 and 0.5), most of the observations for Chat appear in the upper-left quadrant while the observations for Vote appear in the lower left. In other words, it appears that groups are *more likely* to be accurate than inaccurate when they can communicate, while they are *less likely* to be accurate than inaccurate when they cannot. We caution that this finding may not generalize beyond the particular knowledge items in our study, though it is nevertheless quite intriguing. We also note that the effect of communication appears to be widely variable within this range. Indeed there is one item for which Chat increases group accuracy to 100% but another item such that it decreases to 0%![13]

4.2 Beliefs

In this section, we investigate the effects of communication on collective decisions from an informational perspective, analyzing social learning. In doing so, we shift our attention from

[13]Chat increases group accuracy for the factual statement "The first Summer Olympic Games in 1896 were held in Rome, Italy" (which is false, as they were held in Athens, Greece), but decreases accuracy for the statement "Ninety percent of Americans want our background check system strengthened and expanded to cover more gun sales" (PolitiFact rated the statement by Senator Chris Murphy as true, `https://www.politifact.com/truth-o-meter/statements/2016/jul/28/chris-murphy/dnc-sen-chris-murphy-says-90-americans-want-expand/`).

a focus on the aggregation of individual votes to the extent to which individuals update their beliefs based on the opinions of others. We are interested in two questions. First, how much do individuals update their beliefs when they only know how others voted (as in the Vote treatment)? Answering this question will provide an estimate of how informative individuals think others' opinions are. Second, how does communication in the Chat treatment affect how individuals update their beliefs? More specifically, do they update their beliefs in a way that suggests communication provides more information than voting alone?

To ground our analysis, we start by considering a standard Bayesian belief updating model in which the truth of a statement is an unknown binary state of the world, $\omega \in \{T, F\}$. An individual with prior belief π_0 receives a set of K signals $S = (s_1, \ldots, s_K)$, where each signal $s_k \in \{T, F\}$ is interpreted as indicating the truth of the statement. Let the informativeness of signal T be given by $\alpha = \Pr(S_k = T|\omega = T)$ and the informativeness of signal F be given by $\beta = \Pr(S_k = F|\omega = F)$. Upon observing the k-th additional independent signal, a rational individual would update their belief from π_{k-1} to π_k using Bayes' Rule, which we can express in terms of odds ratios according to the formula

$$\frac{\pi_k}{1-\pi_k} = \frac{\pi_{k-1}}{1-\pi_{k-1}} \cdot L_k \,,$$

where L_k is the likelihood ratio of the k-th signal, which is $L_T = \frac{\alpha}{1-\beta}$ for signal T and $L_F = \frac{1-\alpha}{\beta}$ for signal F. If the k signals are independent, then the posterior belief after observing n_T true signals and n_F false signals can be written as a function of the prior belief and the product of the likelihood ratios of the signals,

$$\frac{\pi_k}{1-\pi_k} = \frac{\pi_0}{1-\pi_0} \left(\frac{\alpha}{1-\beta}\right)^{n_T} \left(\frac{1-\alpha}{\beta}\right)^{n_F} .$$

The odds-ratio form of Bayes' Rule can be linearized by taking the log of both sides, yielding

the log-odds form

$$\text{logit}(\pi_k) = \text{logit}(\pi_0) + n_T \log\left(\frac{\alpha}{1-\beta}\right) + n_F \log\left(\frac{1-\alpha}{\beta}\right). \quad (1)$$

The corresponding empirical specification is

$$\text{logit}(\pi_k) = \delta \text{logit}(\pi_0) + \lambda_T n_T + \lambda_F n_F + \varepsilon \quad (2)$$

where δ is the weight on prior beliefs (for a fully Bayesian individual, $\delta = 1$) and λ_T and λ_F are the logs of the likelihood ratios. This specification can be estimated by OLS (without a constant) and is similar to the belief updating models estimated in Mobius et al. (2011), Hill (2017), and Coutts (2019). In those studies, the empirical model can be used to test the extent to which belief updating is Bayesian, as the values of α and β (and hence the likelihood ratios) are known because signals are drawn with objective probabilities controlled by the experimenter (e.g., balls drawn from urns with known distributions). In contrast, the informativeness of signals in our study are not controlled but instead are "home grown" in the sense that our analysis assumes individuals rely on their own beliefs about the informativeness of others' signals. We can then infer properties of these second-order beliefs from the statistical estimates. For example, if both T and F signals from others are believed to be informative, $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$, which implies $\lambda_T > 0$ and $\lambda_F < 0$. If signals are symmetrically informative, then $\alpha = \beta$ implies $\lambda_T = -\lambda_F$.

Table 4 presents estimates of equation (2) separately for the Vote and Chat treatments, with random effects for sessions, and statements included in the model.[14] We use the number of actual votes of other group members (*True votes* and *False votes*) as proxies for the set of signals observed, as in the theoretical framework. This is appropriate in the Vote

[14]Kernel density plots for prior and posterior beliefs aggregated across statements can be found in the Appendix. Posterior beliefs of 0 and 1 cannot be updated in the Bayesian model and, correspondingly, the log odds of such beliefs cannot be computed. Following previous work (Coutts, 2019; Hill, 2017; Mobius et al., 2011), we adjust the endpoints so that extreme beliefs of 0 are replaced with 0.01 and beliefs of 1 are replaced with 0.99.

treatment (column 1), because the number of votes is the only information subjects receive about others when we elicit their posterior beliefs, and so these estimates are the closest empirical analogue to the theoretical setup. The results suggest that subjects act as if the social information they receive is symmetrically informative, as the coefficient for true votes is positive and significant, while the coefficient for false votes is negative and significant, and they are approximately the same magnitude.[15]

If we interpret the coefficients λ_T and λ_F as direct estimates of the likelihood ratios, we can back out estimates of the informativeness of social signals α and β. Since $\lambda_T \approx -\lambda_F$, we can simplify and assume that $\alpha = \beta$ and then compute $\hat{\alpha} = \frac{e^{\hat{\lambda}}}{1+e^{\hat{\lambda}}}$. For the estimate of the likelihood ratio in column 1, $\hat{\lambda} = 0.29$, and this corresponds to an estimate of signal quality $\hat{\alpha} = 0.57$. This suggests that while subjects in the experiment relied on others' votes to update their own beliefs, they did not seem to think that others' information was particularly reliable.

In the Chat treatment, the observed votes are not simply indicators of the independent signals of other group members but instead affected by communication (and therefore correlated). Indeed, Figure 3 shows how communication induces a drastic change in the distribution of votes. Without communication (upper panels), the distributions shown separately for true and false statements are unimodal; with simple majorities, voting true is the mode in both cases. With communication (lower panels), however, simple majorities are rare. Instead, the distributions are bimodal and indicate a preponderance of unanimous votes (65-70% compared to only 15-20% without communication). We note that in Chat, unanimous votes tend to be correct. For true statements (lower left), unanimous votes for true outnumber unanimous votes for false, while for false statements (lower right), unanimous votes for false outweigh unanimous votes for true.

Hence, in this case, we further control for the direction of the exchanged messages[16]

[15]Although we are not really interested in testing whether our subjects are fully Bayesian, we find that individuals underweight their prior beliefs (the coefficient is less than 1), meaning that they update their beliefs as if their prior is closer to $\frac{1}{2}$ than they reported.

[16]The direction of a message was determined by the coding of 2 research assistants, blind to the experi-

Table 4: Belief updating as if others' votes are mildly informative signals

	Dependent variable:			
	logit(posterior)			
	(1)	(2)	(3)	(4)
logit(prior)	0.789	0.574	0.784	0.677
	(0.011)	(0.017)	(0.012)	(0.021)
True votes	0.289	0.229		
	(0.012)	(0.054)		
False votes	−0.293	−0.421		
	(0.013)	(0.052)		
True chats		0.163		
		(0.059)		
False chats		0.016		
		(0.057)		
True beliefs			0.127	−0.075
			(0.019)	(0.048)
False beliefs			−0.136	0.019
			(0.021)	(0.052)
Observations	3,680	2,053	3,680	2,160
Log Likelihood	−6,013.898	−3,668.233	−6,271.832	−4,349.438

*p<0.1; **p<0.05; ***p<0.01

Linear mixed effects model with item, and session random effects. Columns (1) and (3) include observations from the Vote treatment, while (2) and (4) from the Chat one.

by introducing two more variables. The *True chats* and *False chats* variables represent the sum of the other group members' messages in support of the direction True, relative to their total number of messages in support of either direction.[17] The results in Column 2 suggest that belief updating in the Chat treatment differs in two ways from the Vote treatment. First, subjects rely more on social information in Chat than Vote (as the vote coefficients are greater in magnitude) and less on their prior beliefs (with a smaller coefficient). Second,

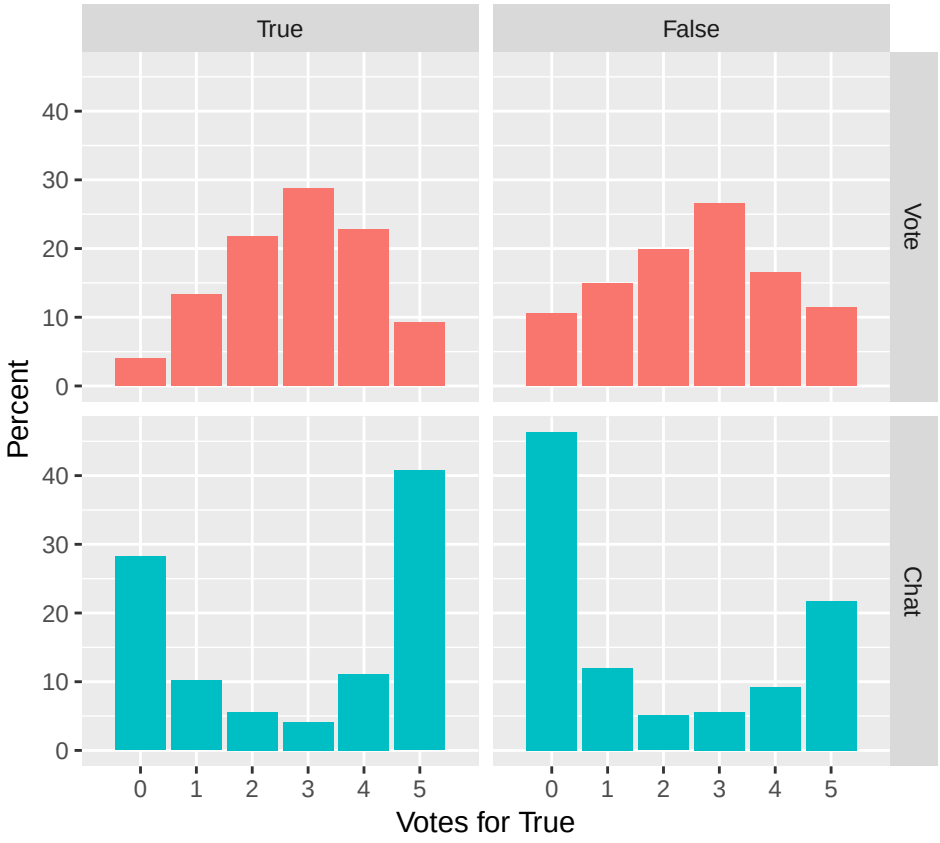


Figure 3: Communication induces unanimous decisions

mental conditions who were asked to specify whether the message is supporting/related to the True or the False direction. Whenever an inter-coder disagreement was present (i.e. one coder considered the message to support the True direction and the other the False direction), we coded that observation as NA. Inter-coder reliability for this measure is very high (Kippendorff's alpha = 0.946).

[17]For each member, we first compute the strength of their support for the True and False directions as expressed in the chat by dividing the number of messages sent by this member in support of the True (False) direction with the sum of messages in support of either direction. This gives us two values between 0 and 1 for each member, which represent the proportion of this member's messages in support of True and in support of False. Afterwards, for each member within each group, we sum over every other member's True (False) chat value to compute their value for the True (False) chats variable.

messages in favor of True seem to have a significant effect on people's beliefs while those in favor of False do not.

We run two alternative models intended to control for the effect of communication in Chat. Specifically, we create alternative measures of others' signals based on their prior beliefs (similar to the approach we took with belief-based accuracy) rather than the observed votes. The variable *true beliefs* is the number of other group members with prior beliefs greater than or equal to $\frac{1}{2}$ and *false beliefs* is the number of group members with prior beliefs less than $\frac{1}{2}$. The belief-based variables are not measures of the signals a subject observes before updating their beliefs, but rather measures of the information possessed by other group members. If this information is transmitted in the same way in the Vote and Chat treatments, then we would expect to see that subjects update their beliefs in the same way in both treatments.

The results in Table 4 (columns 3 and 4) suggest that social information affects beliefs differently in Vote and Chat. For the Vote treatment, the results are qualitatively similar to the specification using actual votes, only the magnitudes for the belief coefficients in column 3 are smaller than the actual vote coefficients in column 1. This is not surprising since observed votes are noisy functions of beliefs (so the coefficients are likely attenuated due to measurement error). For the Chat treatment, as in the column (2) model, we notice that subjects put less weight on their priors compared to Vote. Furthermore, neither belief coefficient is significant—which we find somewhat surprising. These results suggest that groups use communication to do something other than simply transmit and aggregate their information or intended votes.

Result 2. *Participants in the Vote treatment take into account their group members' prior beliefs in their belief updating process. Communication in the Chat treatment diminishes the influence of others' prior beliefs on participants' own posterior beliefs.*

4.3 Content of Communication

How do groups use communication? What do they talk about? To get a better understanding of how communication can affect belief updating, in this section we present a qualitative and quantitative analysis of the chat content.

In some cases, there is no substantive discussion and members simply convey their guess or belief, as in the very brief (yet also complete) transcript shown in Table 5.[18] In other groups, members make factual assertions or claims that, if true, help the group to resolve the question at hand. For example, in the transcript in Table 6 where the fact to be evaluated is whether the first (modern) Olympic Games were held in Rome, the first member to send a message mentions Athens. Two other characteristics of this particular group chat are noteworthy. First, only one member explicitly uses the word "true" (in this case, actually says "isn't true"). Second, members use language that expresses doubt or uncertainty: member 2 says "i don't know" and member 5 says "Yeah I have no clue," and even member 4 (who originally mentioned Athens) then hedges at the end and says "I might be wrong though."

Examples of longer conversations are given in the Appendix (Tables C5 and C6). Examining the transcripts, we see what we might characterize as attempts at interactive reasoning. Such exchanges start with a member initially asserting nothing more than that the statement is true or false, and then other members begin to parse and evaluate the

Table 5: Chat transcript from Session 5, Group 1 discussing Item 12: "The number of unauthorized immigrants to the United States has increased since 2007."

Member	Message
4	i said true
3	i said true
2	i also said true
1	true I think

[18]This statement is false according to estimates from the Pew Research Center: `https://www.pewresearch.org/hispanic/2018/11/27/u-s-unauthorized-immigrant-total-dips-to-lowest-level-in-a-decade/`.

Table 6: Chat transcript from Session 6, Group 2 discussing Item 7: "The first Summer Olympic Games in 1896 were held in Rome, Italy."

Member	Message
4	I thought it was Athens, Greece
2	i don't know the answer to this one
1	i feel like this isn't true but im not sure where it was held
5	Yeah I have no clue
3	Greece seems right
4	I also thought the first Olympics were held before this
4	I might be wrong though

particular details of the statement. In one example, while assessing whether hate speech is protected by the First Amendment (Table C5)[19], there is an exchange between three of the 5 members (1, 4, and 5) in which they discuss the distinction between "free speech" and "hate speech." Similarly, in another example where group members are evaluating a politician's claim about public support for background checks related to gun sales (Table C6)[20], one member asserts that the statistic used ("ninety percent") seems too high, which leads the group to consider what they think the right proportion of the population might be, with one member generalizing from their own family's experience. In these examples, group members make simple arguments, bring their own knowledge to the attention of the group, and discuss the relative merits of those arguments and the credibility of their information. To understand how common it is for groups to engage in processes of interactive reasoning (however rudimentary their attempts might be to a sophisticated observer) we use both a quantitative textual analysis based on the frequency of certain communication measures and a qualitative analysis based on the ratings of two human coders.

Coarse quantitative measures of communication activity support the conclusion that groups do more than communicate their individual judgments, as they engage in simple forms of deliberation and interactive reasoning. Table 7 presents mixed effects models for several

[19]This statement is false according to PolitiFact: https://www.politifact.com/truth-o-meter/statements/2017/apr/21/howard-dean/howard-deans-wrong-tweet-constitution-doesnt-prote/.

[20]This statement is true according to PolitiFact: https://www.politifact.com/wisconsin/statements/2017/oct/03/chris-abele/do-90-americans-support-background-checks-all-gun-/.

Table 7: Chat activity and substance increases with item difficulty

	Dependent variable:			
	Messages	Words	Words/Msg.	%TF Msgs.
	(1)	(2)	(3)	(4)
Medium	1.993	17.352	0.732	−0.100
	(0.912)	(5.926)	(0.251)	(0.024)
Hard	2.062	16.593	0.531	−0.067
	(0.987)	(6.378)	(0.270)	(0.026)
Constant	12.981	63.792	4.619	0.521
	(1.261)	(6.597)	(0.311)	(0.028)
Observations	432	432	432	432
Log Likelihood	−1,497.838	−2,281.206	−842.489	64.615

*p<0.1; **p<0.05; ***p<0.01

Linear mixed effects model with item and session random effects.

measures of activity. Each observation corresponds to one "chat" (a single group discussion of one statement), and each model includes item and session random effects. Each measure of activity is regressed on indicators of medium and hard item difficulty, so the intercept provides an estimate of the mean for easy items. In column 1, we find that groups exchange approximately 13 messages per chat for easy items, which increases to about 15 messages for medium and hard items. Column 2 shows the total number of words used in each chat, which is almost 64 words per chat for easy items, increasing by about 17 words for medium and hard items. In column 3 we consider the average length of a message in each chat, which is about 4.6 words per message for easy items and increases by 0.7 for medium items and 0.5 for hard items. Although these simple measures of activity ignore the substantive content of each message, the increase in activity for more difficult items is consistent with groups using communication for the purposes of interactive reasoning.

As a crude measure of message content, we calculate the percentage of messages in each chat in which subjects use the word "true" or "false" (or some variant such as "t", "f",

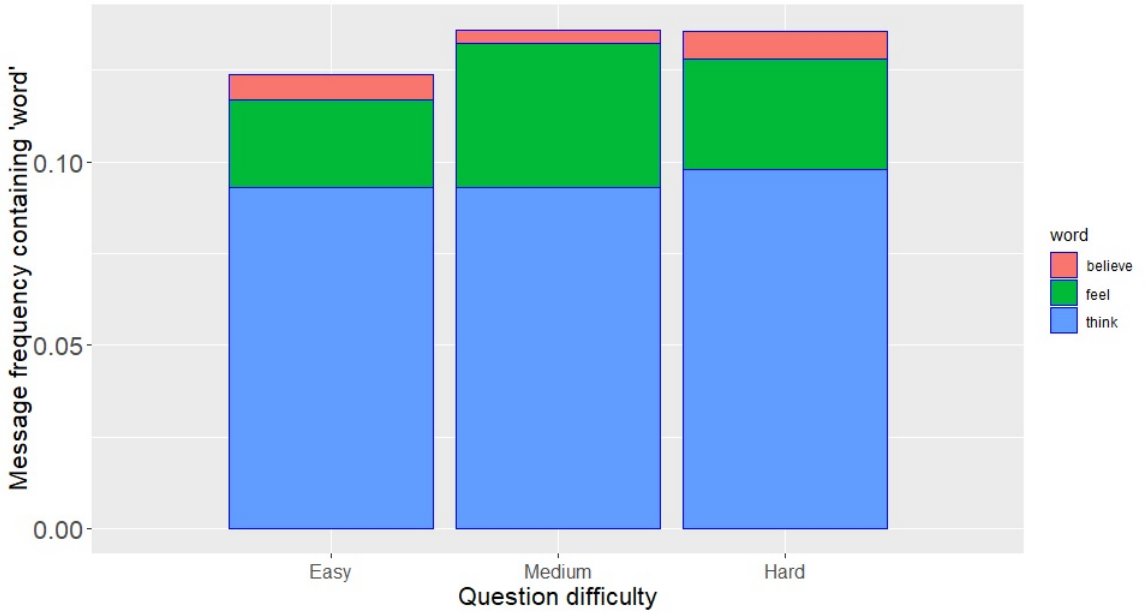
"tru", "fake", "fals"). From column 4 of Table 7, we see that on average 52% of messages for easy items involve the explicit use of the words true or false. For more difficult items, this decreases by roughly 10 percentage points for medium questions and 7 percentage points for hard questions. Thus, we find that our analysis of a simple measure of message content yields similar results as the analysis of chat activity. When items are more difficult, there is an increase in activity associated with greater interactive reasoning and a corresponding decrease in simple declarations of members' guesses.

We also examine two additional measures of content based on the presence of a handful of keywords, abstracting from the specific substantive details of group discussion. As an indicator of information exchange, we look at the frequency of messages containing words such as "think", "feel" or "believe." In many of the messages, these words are associated with tentative assertions of factual claims, such as "I think they had a war in like the early 2000s", "i feel like we made the transition to oil a while agho", or "Murders have been decreasing steadily since the 80s I believe." Panel (a) in Figure 4 suggest that by this measure, about 15% of messages for easy items pertain to the exchange of information, increasing slightly to 17.5% for medium and hard questions.

Next, to obtain a measure associated with reasoning in communication, we identify messages containing words "because", "so", "argument", "but", "maybe", "although", "reason".[21] These are words that are suggestive of a team deliberation process, for example: "i feel like it's true bc the housing market fell in 2008" or "I am not sure because it could still be feeling the effects of the housing crisis or it could be recovering." Panel (b) in Figure 4 depicts the frequency of such messages for the different levels of question difficulty. We notice a sizeable increase in the frequency of interactive reasoning for hard questions (24%) compared to easy questions (16%). Taken together, these results suggest that the mechanism that makes chat more effective for harder questions is a deliberative process that combines the exchange of information with interactive reasoning.

[21] We take into account also variants of these words such as "bc", "cause", "though", etc.

(a) Information exchange



(b) Interactive reasoning

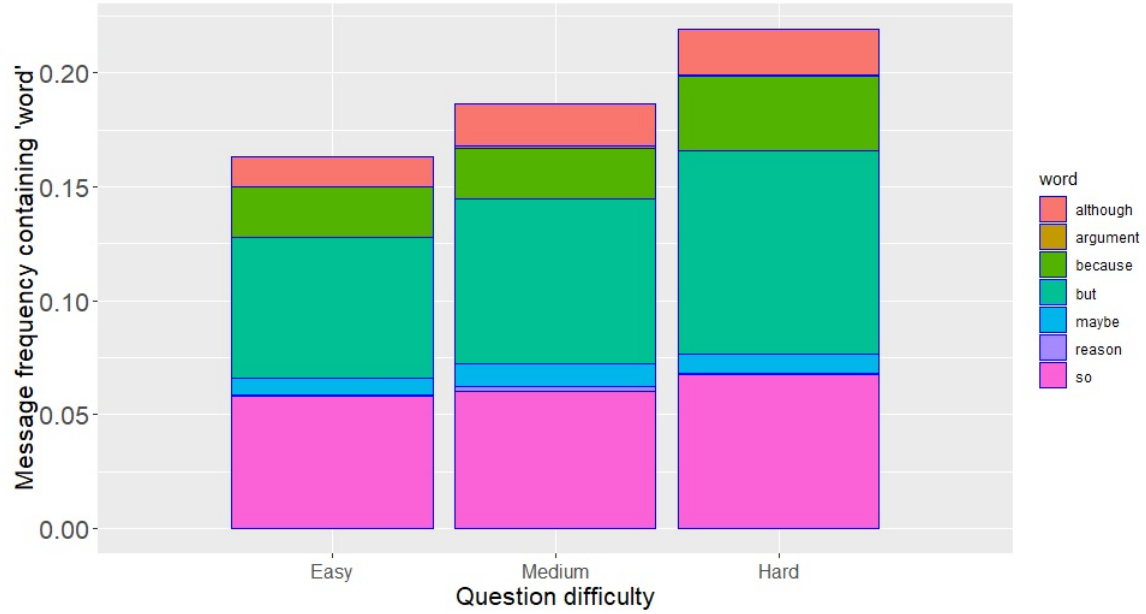


Figure 4: Information exchange and interactive reasoning based on words analysis

We complement this text-based analysis with a human-coded one. Specifically, we asked two research assistants, blind to the experimental conditions, to categorize each message into the following categories: (1) belief (conveys the subject's belief that the statement is true or false); (2) fact (communicates a fact or a supporting piece of information); (3) argument (provides reasons in support of a particular idea or opinion); (4) response (responds to a previous message); (5) doubt (expresses doubt or uncertainty. We find that the coding of the fact variable is positively correlated with that of the argument variable ($\rho = 0.31$, $p < 0.001$) so for each coder, we create another dummy variable that takes the value 1 if the message is coded either as a fact or as an argument or both.[22] For each message and each variable, we take the average across the two coders. All of these measures have either a close to zero (insignificant) correlation or are negatively correlated with each other.[23] Table 8 presents mixed effects models for these measures, with item and session random effects where the unit of observation is one group chat.

First, we notice that there are fewer belief statements the more difficult a question is which is in line with the text-based analysis where such evaluative statements were proxied by the presence of words such as "true" or "false" (Table 7, column 4). The differences in this case is though not significant. Next, as prefaced by the textual analysis in Figure 4, we observe a significant increase in deliberative messages stating facts or arguments for more difficult items (column 2) as well as more interaction (proxied by the amount of messages coded as "responses" to previous statements in the group chat - column 3). It also seems that more difficult items induce higher frequencies of messages expressing doubt (column 4). However, these coefficients are not significant.

[22] Analysis of the inter-coder reliability suggests all our measures are informative. Krippendorff's alpha values are relatively high despite the infrequent nature of some of these types of communication, which biases Krippendorff's alpha values downwards: belief - $\alpha = 0.586$, fact/argument - $\alpha = 0.708$, response - $\alpha = 0.395$, doubt - $\alpha = 0.544$.

[23] For the correlation between belief with fact/argument, $\rho = 0.03$, $p < 0.05$, for belief with response, $\rho = -0.17$, $p < 0.001$ and for belief with doubt, $\rho = -0.04$, $p < 0.01$. For the correlation between fact/argument with response, $\rho = -0.01$, $p > 0.1$ and for fact/argument with doubt, $\rho = -0.02$, $p > 0.1$. Finally, for the correlation between response and doubt, $\rho = -0.21$, $p < 0.001$.

Table 8: Human-coded chat substance

	Dependent variable:			
	Belief	Fact/Argument	Response	Doubt
	(1)	(2)	(3)	(4)
Medium	−0.039	0.092	0.031	0.034
	(0.029)	(0.026)	(0.018)	(0.037)
Hard	−0.026	0.056	0.022	0.018
	(0.031)	(0.028)	(0.019)	(0.039)
Constant	0.616	0.201	0.288	0.226
	(0.027)	(0.026)	(0.022)	(0.035)
Observations	432	432	432	432
Log Likelihood	118.434	213.782	204.329	178.887
Note:			p<0.1; p<0.05;	p<0.01

Result 3. *Groups exchange more and lengthier messages, that include more facts/arguments when items are more di cult.*

4.4 Communication and Group Accuracy

How does group communication affect the accuracy of group decisions? Are groups more accurate if they talk more? Or if they reason more? Do group members who dominate the discussion have more influence over the decision? To answer these questions, we construct measures of the accuracy of key members of each group from the chat transcripts and their prior beliefs and use these measures as independent variables in our analysis along with group-level measures of chat activity and content from the previous section.

We focus on three types of potentially influential members within each group. First, we identify the *most talkative* member(s) of each group in terms of the number of messages sent during the chat period.[24] Second, we identify the *most con dent* member(s) of each group in terms of prior beliefs furthest from $\frac{1}{2}$ for the item under consideration (regardless

[24]Defining the most talkative in terms of the total number of words yields similar results.

of direction). Third, we identify the member who is the *first to speak* within the group chat. Once we identify these members within each group, we then code their belief-based accuracy as before (1 if the belief is in the right direction, 0 if uncertain or in the wrong direction) and, to account for the possibility that there may be more than one most talkative or most confident member of each type, we then take the average of belief-based accuracy. In addition to the accuracy of key members, we also include the number of accurate prior opinions in the group in the analysis.

We also include several measures of chat activity and content based on the entire chat transcript of each group. The two basic measures of chat activity are the total *number of messages* sent and average number of *words per message*. In terms of content, we include the proportion of messages that contain *belief words* (as in Figure 4a), the proportion of messages that contain *deliberation words* (as in Figure 4b), and the proportion of messages that include *truth value* words ("true" or "false").

Table 9 presents the estimates from a linear mixed effects regression with group accuracy as the dependent variable and with session level random effects. Column (1) shows results pooling across item difficulty, while columns (2) through (4) present results for each level of question difficulty separately. Since the dependent variable is dichotomous, the regression is a linear probability model and the coefficients can be interpreted in terms of the change in probability of group accuracy.

The main finding is that certain members of the group appear to be influential, driving the group's decision, while neither the overall levels of chat activity nor the nature of the discussion (as measured via the bag-of-words approach) seem to have much bearing on group accuracy. More specifically, group accuracy is positively and significantly related to both the accuracy of the most talkative and the most confident members of the group. Groups are approximately 18-32 percentage points more accurate when the most talkative member is accurate than when the most talkative member is inaccurate. Similarly, groups are roughly 26-31 percentage points more accurate when the most confident member is accurate rather

Table 9: Group accuracy, influential members, chat activity, and chat content

	Dependent variable: Group accuracy			
	Pooled	Easy	Medium	Hard
	(1)	(2)	(3)	(4)
Accurate priors, most talkative	0.276	0.271	0.319	0.179
	(0.055)	(0.092)	(0.095)	(0.100)
Accurate priors, most confident	0.288	0.312	0.261	0.284
	(0.053)	(0.092)	(0.086)	(0.095)
Accurate priors, first to speak	0.061	−0.159	0.094	0.197
	(0.050)	(0.084)	(0.084)	(0.092)
Number of accurate priors	0.044	0.072	−0.008	0.056
	(0.025)	(0.041)	(0.042)	(0.045)
Number of messages	−0.001	−0.004	−0.002	0.001
	(0.003)	(0.004)	(0.005)	(0.005)
Words per message	−0.016	−0.044	−0.005	0.026
	(0.015)	(0.023)	(0.025)	(0.028)
Prop. belief words	−0.125	−0.171	−0.237	−0.179
	(0.171)	(0.233)	(0.333)	(0.298)
Prop. deliberation words	0.075	0.099	−0.300	0.059
	(0.196)	(0.303)	(0.357)	(0.349)
Prop. truth value words	−0.138	−0.162	0.047	−0.169
	(0.119)	(0.163)	(0.230)	(0.216)
Medium difficulty	−0.024			
	(0.055)			
Hard difficulty	0.016			
	(0.067)			
Constant	0.323	0.556	0.383	0.091
	(0.130)	(0.191)	(0.204)	(0.197)
Observations	432	126	165	141
Log Likelihood	−240.249	−44.685	−111.367	−89.765

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Linear mixed effects model with session random effects.

than inaccurate. Whether the accuracy of the group depends on the first member to speak depends on item difficulty. For hard items, groups are estimated to be 20 percentage points more accurate when the first speaker is accurate, but 16 percentage points less accurate for easy items (significant at the 0.10 level), with a positive but insignificant coefficient for medium items.

Turning to the coefficients for chat activity and content, we find that most of the coefficients are not statistically significant. For easy items, the number of words per message is negatively associated with accuracy (though significant at the .10 level). The coefficient for the proportion of words associated with beliefs are consistently negative and between 0.17 and 0.24 in magnitude, though none of these coefficients are significant. Similarly, the coefficient for deliberation words is not significant. Taken together, the results for belief and deliberation words is mildly suggestive that groups that express greater uncertainty or doubt via more deliberation or possibly opposing belief statements are unable to resolve that uncertainty in favor of accuracy.

Next, we note that the accuracy of the group decision is increasing in the accuracy of group members' prior opinions only for easy items (significant at the .10 level), but not for medium or hard items. This is consistent with what we found in our analysis of individual belief updating (section 4.2) and with the implication that group discussion involves something other than simple exchange of information or beliefs.

The coarse bag-of-words based measures for chat content fail to capture the contextual meaning of the messages. To account for this, we run a similar regression using the human-coded measures for identifying the proportion of expressed beliefs and proportion of deliberation in the chats (as measured by the frequency of messages stating a fact or an argument or responding to a previous message). Table 10 presents the estimates from a linear mixed effects regression with group accuracy as the dependent variable and with session level random effects. Instead of the proportion of belief or truth value words, we use the proportion of messages coded as expressing the participant's belief about the truth

value of the statement. Next, the proportion of deliberation words is replaced with two other variables: one measuring the proportion of messages stating facts or making arguments in support or against the particular item, and one measuring the proportion of messages in the chat that are written in response to each other (this is a rough proxy for the extent to which individuals engage with other group members' statements rather than simply expressing their independent view).

The results are broadly in line with the bag-of-words based analysis. We again find that the most talkative and the most confident members are those that, irrespective of the item difficulty, significantly and positively contribute to the group's accuracy. The coefficient for the first one to speak has the same sign as before (negative for easy items and positive for medium and hard ones) as well as similar magnitude. The number of members with accurate priors does not seem to have an impact on group accuracy (not even for easy items). A higher chat volume, as measured by a higher number of messages, reduces the group accuracy but only slightly (with 1.6 percentage points) and only for medium difficulty items, while a higher number of words per message does not have a significant effect.

The coefficients for the variables measuring the effect of the chat content as coded by human coders, seem to be more informative than the proxies using the bag-of-words approach. In particular, we find that a higher proportion of messages including statements of belief leads to a significant decrease in group accuracy of 29 percentage points across all item difficulties. When looking at the coefficients in columns 2-4, we see that this effect is mostly driven by the medium difficulty items - the only significant coefficient across the three models and the highest in magnitude. Turning to the deliberation variables, we find that the proportion of facts or arguments individuals bring to the discussion does not have a significant effect for either difficulty level. However, the proportion of interaction, proxied by the proportion of messages coded as being a response to previous messages in the group chat, seems to have a significantly negative and large (47 percentage points) effect on group accuracy for easy items but positive for more difficult items. However, only the coefficient

Table 10: Group accuracy, influential members, chat activity, and human-coded chat content

	Dependent variable: Group accuracy			
	Pooled	Easy	Medium	Hard
	(1)	(2)	(3)	(4)
Accurate priors, most talkative	0.277	0.270	0.302	0.204
	(0.055)	(0.090)	(0.089)	(0.101)
Accurate priors, most confident	0.271	0.332	0.205	0.310
	(0.053)	(0.090)	(0.085)	(0.093)
Accurate priors, first to speak	0.053	−0.167	0.119	0.162
	(0.050)	(0.083)	(0.079)	(0.094)
Number of accurate priors	0.046	0.061	0.001	0.065
	(0.024)	(0.040)	(0.040)	(0.045)
Number of messages	−0.004	0.001	−0.016	0.002
	(0.003)	(0.004)	(0.005)	(0.006)
Words per message	0.0002	−0.036	0.007	0.023
	(0.015)	(0.025)	(0.023)	(0.029)
Prop. belief	−0.292	−0.027	−0.716	0.027
	(0.136)	(0.190)	(0.237)	(0.260)
Prop. facts/arguments	−0.262	−0.086	−0.391	−0.129
	(0.178)	(0.299)	(0.268)	(0.363)
Prop. response	0.195	−0.466	0.654	0.358
	(0.138)	(0.189)	(0.233)	(0.282)
Medium difficulty	−0.015			
	(0.055)			
Hard difficulty	0.018			
	(0.067)			
Constant	0.402	0.573	0.772	−0.112
	(0.157)	(0.230)	(0.232)	(0.273)
Observations	432	126	165	141
Log Likelihood	−235.835	−42.339	−101.606	−89.363

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Linear mixed effects model with session random effects.

for medium items is significant. This suggests that when facing easier items, group members are more likely to confuse each other the more they interactively reason, whereas for more difficult items, a higher proportion of interactive reasoning allows them to more efficiently bridge their individual gaps of knowledge.

Result 4. *When communication is allowed, group accuracy depends on the accuracy of the most talkative and the most confident member, irrespective of the item difficulty. A higher volume of communication has a small effect on group accuracy only for medium difficulty items. A higher proportion of interactive communication (response messages) harms group accuracy for easy items and helps it for more difficult ones.*

5 Conclusion

We find that communication before voting generally helps groups to figure out the truth, getting it right more often than they do when they decide by majority rule without discussion. However, group performance is better mainly for hard items and worse for easy ones. This shows that even when communication is cheap, and there is no cost to ignoring it, group decisions are influenced by the costless exchange of subjective statements. Moreover, we identify an important moderator for the effect of communication on group performance: item difficulty. Morton, Piovesan and Tyran (2019) also found that the effect of social information on group performance depends on item difficulty. However, in their study, where groups are presented with non-interactive social information (about how other group members voted), the opposite dynamic is observed: social information harms when the task is difficult, and helps when the task is easy.

The chat analysis points to one possible reason why communication harms group accuracy for easy items and improves it for more difficult ones in our setting. Specifically, we find that chat activity and content are consistent with groups using communication for information exchange and interactive reasoning. Such interactive reasoning may induce more

uncertainty and potential confusion when the item is easy and there are not many gaps in individuals' knowledge to be complemented. However, when the item is more difficult, and each individual's knowledge is sparse on the topic but each member may have different pieces of information, interactive reasoning helps groups deal with the existing uncertainty and bring the members' knowledge together in an efficient manner.

This evidence suggests that collective decisions may still benefit from social information, across all levels of item difficulty, provided that the right format is present. Whether groups are able to choose the optimal type of social information could represent an interesting avenue for future research.

We also find that chat improves individual accuracy for all levels of question difficulty. This is at odds with previous studies investigating the effect of non-interactive social information such as Lorenz et al. (2011), Novaes Tump et al. (2018) and He, Lien and Zheng (2021) which suggest that this type of social information helps individual performance for easy items and harms it for difficult ones. This underscores the importance of further investigating how different interactive or non-interactive social information affects individual judgments. This is relevant not only in collective judgment tasks, but also in individual ones as previous research has found that non-interactive social information reduces the average individual accuracy when people are rewarded for their individual responses but increases it when rewarded for the collective performance (Bazazi et al., 2019).

Finally, our study shows that in the absence of communication, individual beliefs change in ways consistent with the reliance on others' votes as mildly informative signals. However, when communication is present, people put less weight on their priors and more weight on the social, interactive, information. Moreover, the volume of information exchange and interactive reasoning increases with the item difficulty. Our results also reveal that across all levels of difficulty, group communication allows certain members of the group, i.e. the most talkative and most confident ones, to influence to a great extent the group's accuracy. Our study can provide the framework for investigating whether the same dynamics are

present in larger groups or when the collective decision is made using rules different than majority voting.

References

Allen, Franklin. 1987. "Discovering personal probabilities when utility functions are unknown." *Management Science* 33(4):542–544.

Austen-Smith, David and Jeffrey S Banks. 1996. "Information aggregation, rationality, and the Condorcet jury theorem." *American Political Science Review* 90(1):34–45.

Bartels, Larry M. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.

Bazazi, Sepideh, Jorina von Zimmermann, Bahador Bahrami and Daniel Richardson. 2019. "Self-serving incentives impair collective decisions by increasing conformity." *PloS one* 14(11):e0224725.

Berinsky, Adam J. 2015. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2):241–262.

Bullock, John G, Alan S Gerber, Seth J Hill and Gregory A Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10:519–578.

Cooper, David J and John H Kagel. 2005. "Are two heads better than one? Team versus individual play in signaling games." *American Economic Review* 95(3):477–509.

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." *Experimental Economics* 22(2):369–395.

Delli Carpini, Michael X. and Scott Keeter. 1996. *What Americans know about Politics and Why it Matters*. Yale University Press.

DeMarzo, Peter M, Dimitri Vayanos and Jeffrey Zwiebel. 2003. "Persuasion bias, social influence, and unidimensional opinions." *Quarterly Journal of Economics* 118(3):909–968.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10(2):171–178.

Gaines, Brian J, James H Kuklinski, Paul J Quirk, Buddy Peyton and Jay Verkuilen. 2007. "Same facts, different interpretations: Partisan motivation and opinion on Iraq." *Journal of Politics* 69(4):957–974.

Galton, Francis. 1907. "Vox Populi." *Nature* 750:450–451.

Gilens, Martin. 2001. "Political ignorance and collective policy preferences." *American Political Science Review* 95(2):379–396.

Goeree, Jacob K and Leeat Yariv. 2011. "An experimental study of collective deliberation." *Econometrica* 79(3):893–921.

Grether, David M. 1992. "Testing Bayes rule and the representativeness heuristic: Some experimental evidence." *Journal of Economic Behavior & Organization* 17(1):31–57.

He, Yunwen, Jaimie W Lien and Jie Zheng. 2021. "Stuck in the Wisdom of Crowds: Information, Knowledge, and Heuristics." *Working paper* .

Hill, Seth J. 2017. "Learning together slowly: Bayesian learning about political facts." *Journal of Politics* 79(4):1403–1418.

Holt, Charles A and Angela M Smith. 2016. "Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes." *American Economic Journal: Microeconomics* 8(1):110–39.

Janis, Irving Lester. 1982. *Groupthink: Psychological studies of policy decisions and ascoes*. Vol. 349 Houghton Mifflin Boston.

Jasny, Lorien, Joseph Waggle and Dana R Fisher. 2015. "An empirical examination of echo chambers in US climate policy networks." *Nature Climate Change* 5(8):782.

Jerit, Jennifer and Jason Barabas. 2012. "Partisan Perceptual Bias and the Information Environment." *Journal of Politics* 74(03):672–684.

Karni, Edi. 2009. "A mechanism for eliciting probabilities." *Econometrica* 77(2):603–606.

Kelley, Truman Lee. 1925. "The applicability of the Spearman-Brown formula for the measurement of reliability." *Journal of Educational Psychology* 16(5):300.

Kugler, Tamar, Edgar E Kausel and Martin G Kocher. 2012. "Are groups more rational than individuals? A review of interactive decision making in groups." *Wiley Interdisciplinary Reviews: Cognitive Science* 3(4):471–482.

Kuklinski, J H, P J Quirk, J Jerit and D Schwieder. 2000. "Misinformation and the currency of democratic citizenship." *Journal of Politics* .

Landemore, Hélène. 2017. *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton University Press.

List, Christian and Robert E Goodin. 2001. "Epistemic democracy: Generalizing the Condorcet jury theorem." *Journal of Political Philosophy* 9(3):277–306.

Lorenz, Jan, Heiko Rauhut and Bernhard Kittel. 2015. "Majoritarian democracy undermines truth-finding in deliberative committees." *Research & Politics* 2(2):2053168015582287.

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer and Dirk Helbing. 2011. "How social influence can undermine the wisdom of crowd effect." *Proceedings of the National Academy of Sciences* 108(22):9020–9025.

Lupia, Arthur and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge: Cambridge University Press.

Mercier, Hugo and Dan Sperber. 2017. *The enigma of reason*. Harvard University Press.

Mercier, Hugo and Hélène Landemore. 2012. "Reasoning is for arguing: Understanding the successes and failures of deliberation." *Political psychology* 33(2):243–258.

Mercier, Hugo and Nicolas Claidière. 2022. "Does discussion make crowds any wiser?" *Cognition* 222:104912.

Mobius, Markus M, Muriel Niederle, Paul Niehaus and Tanya S Rosenblat. 2011. Managing self-confidence: Theory and experimental evidence. Technical report National Bureau of Economic Research.

Morton, Rebecca B, Marco Piovesan and Jean-Robert Tyran. 2019. "The dark side of the vote: Biased voters, social information, and information aggregation through majority voting." *Games and Economic Behavior* 113:461–481.

Navajas, Joaquin, Tamara Niella, Gerry Garbulsky, Bahador Bahrami and Mariano Sigman. 2018. "Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds." *Nature Human Behaviour* 2(2):126.

Novaes Tump, Alan, Max Wolf, Jens Krause and Ralf HJM Kurvers. 2018. "Individuals fail to reap the collective benefits of diversity because of over-reliance on personal information." *Journal of the Royal Society Interface* 15(142):20180155.

Robbett, Andrea and Peter Hans Matthews. 2018. "Partisan bias and expressive voting." *Journal of Public Economics* 157:107–120.

Stroop, John Ridley. 1932. "Is the judgment of the group better than that of the average member of the group?" *Journal of Experimental Psychology* 15(5):550.

Surowiecki, James. 2004. *The Wisdom of Crowds*. Anchor.

Woolley, Anita Williams, Christopher F Chabris, Alex Pentland, Nada Hashmi and Thomas W Malone. 2010. "Evidence for a collective intelligence factor in the performance of human groups." *Science* 330(6004):686–688.

Woon, Jonathan. 2019. "Political Lie Detection." Working paper: University of Pittsburgh.

Woon, Jonathan and Kristin Kanthak. 2019. "Elections, ability, and candidate honesty." *Journal of Economic Behavior & Organization* 157:735–753.

Getting it Right: Communication, Voting, and Collective Truth Finding

Valeria Burdea Jonathan Woon

Online Appendix

A Additional experimental details

A.1 Statements

Table A1: Statements used in the experiment

Priors and posteriors are given as the mean beliefs the statements are true; group decisions are proportions of groups with majorities voting true.

	Set	Statement	T/F	Prior (T)		Posterior (T)		Group (T)	
				V	C	V	C	V	C
1	F	Since President Trump took office in 2017, the civilian unemployment rate has decreased by almost 1 percentage point.	True	0.66	0.56	0.74	0.69	0.96	0.75
2	F	The total public debt of the United States federal government more than doubled from quarter 2 in 1981 to quarter 1 in 1989 while Ronald Reagan was president.	True	0.61	0.62	0.67	0.64	0.83	0.75
3	F	From January 2001, when President Bush first took office, to January 2005, when President Bush started his second term in office, the civilian unemployment rate increased by more than 1 percentage point.	True	0.64	0.62	0.71	0.60	0.91	0.58
4	F	The United Kingdom contributes more to the NATO budget than the United States.	False	0.51	0.48	0.54	0.33	0.65	0.25
5	F	There were more gun-related suicides than homicides in the United States in 2016.	True	0.54	0.56	0.53	0.59	0.48	0.67
6	F	From 2009, when President Obama took office, to 2012, median household income adjusted for inflation in the United States fell by more than 4 percent.	True	0.55	0.47	0.66	0.48	0.87	0.50
7	F	The first Summer Olympic Games in 1896 were held in Rome, Italy.	False	0.42	0.48	0.36	0.18	0.35	0.00
8	F	The Battle of Waterloo was fought in Belgium.	True	0.44	0.38	0.43	0.43	0.35	0.33
9	F	The world's busiest airport by passenger traffic is Hartsfield-Jackson Atlanta International Airport.	True	0.50	0.56	0.48	0.57	0.35	0.58

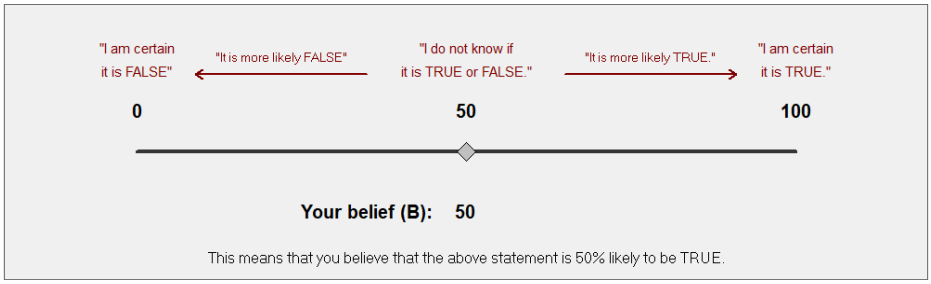
10	F	The rate at which American women aged 15-44 had legal abortions fell more between 1980 and 1988, while Ronald Reagan was president, than between 1992 and 2000, while Bill Clinton was president.	False	0.58	0.52	0.59	0.48	0.70	0.50
11	F	More people in the United States work in the coal industry than in the solar industry.	False	0.63	0.57	0.70	0.51	0.87	0.42
12	F	The number of unauthorized immigrants to the United States has increased since 2007.	False	0.70	0.70	0.77	0.77	0.91	0.75
13	F	The United States spent more on Social Security payments in 2018 than it did on national defense.	True	0.33	0.34	0.28	0.34	0.09	0.25
14	F	West Virginia was part of the Confederacy during the American Civil War.	False	0.57	0.54	0.57	0.31	0.61	0.17
15	F	From 2009 to 2016 while President Obama was in office, the national murder rate decreased by more than 1 percentage point.	False	0.57	0.64	0.62	0.69	0.78	0.75
16	F	There were more refugees to the United States in 2017 from Syria than from Congo.	False	0.70	0.67	0.77	0.69	1.00	0.75
17	P	Hedge fund managers pay less in taxes than nurses and truck drivers.	False	0.57	0.56	0.57	0.48	0.61	0.40
18	P	More than 64 percent of minimum-wage earners are women.	True	0.65	0.65	0.70	0.71	0.91	0.80
19	P	When the US got the income tax in 1913, the top rate was 7 percent. By 1980, the top rate was 70 percent.	True	0.49	0.50	0.46	0.56	0.39	0.60
20	P	Foreign aid is less than 1 percent of our federal budget.	True	0.52	0.54	0.55	0.53	0.57	0.40
21	P	Ninety percent of Americans want our background check system strengthened and expanded to cover more gun sales.	True	0.52	0.47	0.46	0.24	0.35	0.00
22	P	African-American children are 500 percent more likely to die from asthma than white kids.	True	0.46	0.48	0.48	0.42	0.43	0.33
23	P	We are now, for the first time ever, energy independent.	False	0.25	0.22	0.18	0.11	0.00	0.00
24	P	Hate speech is not protected by the first amendment.	False	0.53	0.56	0.49	0.38	0.48	0.33
25	P	We spend almost twice as much per capita on health care as do the people of any other country.	False	0.65	0.69	0.69	0.66	0.91	0.67
26	P	Carbon dioxide is not a primary contributor to the global warming that we see.	False	0.22	0.17	0.17	0.14	0.00	0.00

27	P	Every time we raise the minimum wage, the number of jobless people increases.	False	0.49	0.49	0.47	0.59	0.52	0.67
28	P	Since 2011, the gap in high school graduation rates between African-American students and all students has been cut in half, from 6.4 percent to 3.1 percent.	True	0.60	0.62	0.65	0.61	0.87	0.67
29	P	We have lost more lives in the last two years due to opioids than all of the lives lost during the Vietnam War.	True	0.61	0.68	0.68	0.78	0.87	0.93
30	P	One out of every five people that the federal government charges with murder is an illegal alien.	False	0.28	0.30	0.20	0.24	0.04	0.13
31	P	Over the last eight years, we've shown that even though we've cut taxes by $8 billion, revenues continue to grow.	True	0.52	0.57	0.53	0.68	0.52	0.80
32	P	At $93 trillion, the Green New Deal would cost more than the entire recorded spending of the U.S. since the Constitution went into effect in 1789.	False	0.48	0.44	0.39	0.34	0.30	0.20

A.2 Belief elicitation

A.2.1 Belief elicitation summary provided to subjects

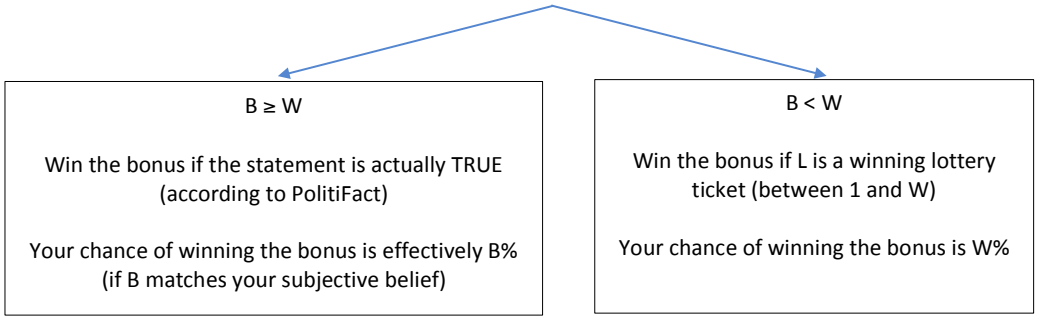
Step 1. Report your belief B that the statement is TRUE by dragging a slider



100	Certain the statement is TRUE
51-99	Statement likely to be TRUE, higher numbers indicate greater certainty it is TRUE
50	Totally uncertain
1-49	Statement likely to be FALSE, lower numbers indicate greater certainty it is FALSE
0	Certain the statement is FALSE

Step 2. We randomly select the number of winning tickets W (numbers 1 to W out of 100)

Step 3. We compare your belief B to W, which affects how the bonus is determined



Remember: Regardless of what belief you hold, you will always have the greatest chance of winning the bonus if you choose B to match your actual belief.

- When your belief gives a higher chance of winning (than the lottery), we pay you for your belief.
- When the lottery gives a higher chance of winning (than your belief), we pay you for the lottery.

The best thing to do is to choose B to match your belief that the statement is TRUE.

A.2.2 Measures of understanding of the belief elicitation method

We used two measures to test subjects' understanding of the belief elicitation method: (i) multiple-choice questions regarding the instructions, and (ii) calibration questions in which subjects had to report beliefs about abstract events.

(i) We used 6 control questions which we list below:

1. If you are certain a statement is TRUE and you report a belief (B) equal to 100, i.e. you think the statement is 100% likely to be TRUE, then:
   - You win the bonus if the statement is indeed TRUE.
   - You will be entered into a lottery if the statement is indeed TRUE.
   - You may be entered into a lottery, depending on the value of the randomly drawn number W.
2. Suppose you state a belief (B) equal to 28, i.e. you think the statement is 28% likely to be TRUE, then:
   - You win the bonus only if the statement is FALSE.
   - You will be entered into a lottery if the statement is indeed TRUE.
   - You may be entered into a lottery, depending on the value of the randomly drawn number W.
3. Suppose you state a belief (B) equal to 44, i.e. you think the statement is 44% likely to be TRUE, and the randomly drawn value of W is 26 (there are 26 winning tickets) then:
   - You win the bonus only if the statement is indeed TRUE.
   - You will be entered into a lottery, irrespective of the truth of the statement.
   - You will be entered into a lottery if the statement is FALSE.
4. Suppose you state a belief (B) equal to 56, i.e. you think the statement is 56% likely to be TRUE, and the randomly drawn value of W is 72 (there are 72 winning tickets) then:
   - You win the bonus only if the statement is indeed TRUE.
   - You will be entered into a lottery, irrespective of the truth of the statement.
   - You will be entered into a lottery if the statement is FALSE.
5. Suppose you state a belief (B) equal to 67 (i.e. you think the statement is 67% likely to be TRUE) and the randomly drawn value of W is 82. How likely is it that you will win the bonus?
   - 82%
   - It depends on whether the statement is TRUE or FALSE.
   - 70%
6. Suppose your best guess is that a given statement is 42 percent likely to be TRUE. Which belief should you state in order to maximize your earnings?

- 100
- 53
- 42

We then count how many times each subject responded erroneously to each question. Summary statistics for each question as well as for the total number of control questions errors are presented in Table A3. We also compare if the average number of total errors in control questions differs across treatments. Using two-sided Wilcoxon rank-sum tests, we find no significant differences across any two treatments (see Table A2).

Table A2: Wilcoxon rank-sum test for pairwise treatment comparison of average total errors in control questions

	p-value
chat-F vs chat-P	0.152
chat-F vs vote-FP	0.743
chat-F vs vote-PF	0.524
chat-P vs vote-PF	0.450
chat-P vs vote-FP	0.323
vote-PF vs vote-FP	0.820

Table A3: Summary statistics for errors in control questions across treatments

		chat-F	chat-P	vote-FP	vote-PF	pooled
ctrl_error_1						
	mean	0.517	0.267	0.317	0.364	0.36
	sd	0.833	0.553	0.676	0.589	0.669
	min	0	0	0	0	0
	max	3	2	2	2	3
ctrl_error_2						
	mean	0.283	0.227	0.350	0.291	0.284
	sd	0.524	0.535	0.577	0.533	0.541
	min	0	0	0	0	0
	max	2	2	2	2	2
ctrl_error_3						
	mean	0.800	0.600	0.733	0.800	0.724
	sd	0.935	0.788	0.880	0.826	0.855
	min	0	0	0	0	0
	max	2	2	3	2	3
ctrl_error_4						
	mean	0.567	0.427	0.550	0.364	0.476
	sd	0.890	0.720	0.852	0.677	0.787
	min	0	0	0	0	0
	max	3	2	4	2	4
ctrl_error_5						
	mean	0.300	0.320	0.400	0.309	0.332
	sd	0.561	0.498	0.588	0.540	0.543
	min	0	0	0	0	0
	max	2	2	2	2	2
ctrl_error_6						
	mean	0.133	0.067	0.100	0.091	0.096
	sd	0.430	0.300	0.354	0.290	0.345
	min	0	0	0	0	0
	max	2	2	2	1	2
ctrl_error_total						
	mean	2.600	1.907	2.450	2.218	2.272
	sd	2.423	1.702	2.310	1.931	2.098
	min	0	0	0	0	0
	max	12	6	8	8	12

(ii) We used 3 calibration items. For each of these items, we asked subjects to report their belief that the statement is true. After each item, participants received full feedback regarding the random draws, whether they were paid according to their belief or according to the lottery, and whether they won the bonus. For each item, subjects were told that after they report the belief, they will roll a computerized fair 6-sided die where each side is numbered from 1 to 6. The three items are listed below. Each subject was presented with these items in a random order.

1. The outcome of the die roll is less than or equal to 6.
2. The outcome of the die roll is 8.
3. The outcome of the die roll is 1, 2 or 3.

The correct answer for item 1. is a belief equal to 100, for item 2. it is equal to 0, whereas for item 3. subjects should report a belief of 50. Table A4 presents summary statistics for the errors subjects made when reporting these beliefs, while Table A5 presents the test results from pairwise treatment comparisons of the average total errors. The average number of errors in reporting these beliefs is relatively low and there is no significant difference across treatments.

Table A4: Summary statistics for errors in control questions across treatments

		chat-F	chat-P	vote-FP	vote-PF	pooled
error_b1						
	mean	0.150	0.067	0.100	0.073	0.096
	sd	0.360	0.251	0.303	0.262	0.295
	min	0	0	0	0	0
	max	1	1	1	1	1
error_b2						
	mean	0.117	0.067	0.150	0.109	0.108
	sd	0.324	0.251	0.360	0.315	0.311
	min	0	0	0	0	0
	max	1	1	1	1	1
error_b3						
	mean	0.283	0.280	0.233	0.345	0.284
	sd	0.454	0.452	0.427	0.480	0.452
	min	0	0	0	0	0
	max	1	1	1	1	1
error_b_total						
	mean	0.550	0.413	0.483	0.527	0.488
	sd	0.872	0.617	0.833	0.766	0.767
	min	0	0	0	0	0
	max	3	2	3	3	3

Table A5: Wilcoxon rank-sum test for pairwise treatment comparison of average total errors in calibration items

	p-value
chat-F vs chat-P	0.678
chat-F vs vote-FP	0.724
chat-F vs vote-PF	0.870
chat-P vs vote-PF	0.530
chat-P vs vote-FP	0.981
vote-PF vs vote-FP	0.572

B Sample characteristics

Table B1: Summary statistics for sample characteristics

	Chat (N = 135)	Vote (N = 115)
Age		
mean (sd)	20.26 ± 1.82	20.40 ± 2.03
Sex		
Female	65%	66%
Male	33%	33%
Other	1%	1%
Race / Ethnicity		
African American	9%	10%
Asian	25%	21%
Latino	2%	2%
White	65%	68%
Major		
Arts and Humanities	14%	14%
Biological Sciences	28%	22%
Physical Sciences	13%	9%
Social Sciences	14%	20%
Business	0%	3%
Other	31%	32%
Party Affiliation		
Democrat	53%	61%
Independent	25%	17%
Republican	15%	16%
Other	7%	7%

C Additional tables and figures

C.1 Separate analysis for P and F statements

Table C1: Individual and group accuracy by treatment and item difficulty: P statements

Accuracy Measure	Easy		Medium		Hard		Total	
	Vote	Chat	Vote	Chat	Vote	Chat	Vote	Chat
Prior beliefs	0.68	0.71	0.50	0.51	0.42	0.42	0.56	0.57
Posterior beliefs	0.74	0.77	0.52	0.51	0.40	0.47	0.58	0.60
Individual votes	0.78	0.86	0.51	0.50	0.39	0.50	0.59	0.64
Group decisions	0.93	0.88	0.51	0.48	0.29	0.51	0.63	0.63
N beliefs/votes	690	450	805	525	345	225	1840	1200
N groups	138	90	161	105	69	45	368	240

Table C2: Individual and group accuracy by treatment and item difficulty: F statements

Accuracy Measure	Easy		Medium		Hard		Total	
	Vote	Chat	Vote	Chat	Vote	Chat	Vote	Chat
Prior beliefs	0.63	0.60	0.53	0.53	0.38	0.39	0.47	0.47
Posterior beliefs	0.70	0.65	0.55	0.63	0.34	0.41	0.47	0.52
Individual votes	0.74	0.71	0.52	0.69	0.30	0.41	0.45	0.55
Group decisions	0.90	0.69	0.54	0.70	0.20	0.41	0.43	0.55
N beliefs/votes	345	180	575	300	920	480	1840	960
N groups	69	36	115	60	184	96	368	192

Table C3: Effect of chat depends on item difficulty. P statements

	Dependent variable:			
	Prior Belief	Posterior Belief	Individual Vote	Group Decision
	(1)	(2)	(3)	(4)
Chat	0.007	−0.016	−0.008	−0.032
	(0.016)	(0.025)	(0.033)	(0.056)
Chat × Easy	0.019	0.040	0.082	−0.024
	(0.022)	(0.023)	(0.037)	(0.076)
Chat × Hard	−0.015	0.088	0.117	0.254
	(0.027)	(0.028)	(0.046)	(0.094)
Easy	0.182	0.219	0.271	0.425
	(0.035)	(0.049)	(0.065)	(0.085)
Hard	−0.079	−0.121	−0.118	−0.219
	(0.043)	(0.061)	(0.080)	(0.106)
Constant	0.503	0.524	0.512	0.509
	(0.024)	(0.036)	(0.046)	(0.060)
Observations	3,040	3,040	3,040	608
Log Likelihood	−345.964	−471.095	−1,915.851	−342.858

*p<0.1; **p<0.05; ***p<0.01

Linear mixed effects model with subject, item, and session random effects in columns 1-3, and item and session random effects in column 4.

Table C4: Effect of chat depends on item difficulty.F statements

	Dependent variable:			
	Prior Belief	Posterior Belief	Individual Vote	Group Decision
	(1)	(2)	(3)	(4)
Chat	−0.008	0.072	0.165	0.161
	(0.021)	(0.026)	(0.040)	(0.068)
Chat × Easy	−0.028	−0.131	−0.201	−0.365
	(0.030)	(0.033)	(0.054)	(0.111)
Chat × Hard	0.015	0.005	−0.063	0.050
	(0.024)	(0.026)	(0.042)	(0.086)
Easy	0.101	0.150	0.220	0.359
	(0.034)	(0.063)	(0.074)	(0.110)
Hard	−0.155	−0.217	−0.217	−0.343
	(0.027)	(0.049)	(0.058)	(0.086)
Constant	0.533	0.555	0.522	0.539
	(0.022)	(0.040)	(0.047)	(0.067)
Observations	2,800	2,800	2,800	560
Log Likelihood	−270.596	−487.990	−1,850.086	−335.280

*p<0.1; **p<0.05; ***p<0.01

Linear mixed effects model with subject, item, and session random effects in columns 1-3, and item and session random effects in column 4.

C.2 Chat transcript examples

Table C5: Chat transcript from Session 2, Group 4 discussing Item 24: "Hate speech is not protected by the first amendment."

Member	Message
5	true
4	it isn
4	true
2	true by guess
1	the first amendment protects freedom of speech though
1	Idk if hate speech is free speech
4	that
5	yeah but hate speech is very specific. I don't think there's anything specifically about hate speech
4	is targetting of people. free speech is more crtical of government and authority wont land you in trouble
1	you right
1	Can you get arrested for hate speech?
5	I don't think so
5	yall saw that video on insta?
1	yes smh
5	it was hate speech but they couldnt do anything
4	but someone has to prosecute?
1	aer we going t or f
2	T
1	are lol
5	I'm going true
1	i will too

Table C6: Chat transcript for Session 12, Group 4 discussing Item 12: "Ninety percent of Americans want our background check system strengthened and expanded to cover more gun sales."

Member	Message
3	false
1	90 is too high
2	i feel like ive heard this before tho
2	really?
1	I think theres alot on the right side who dont agree with this
5	background checks are good tho. why would they disagree?
3	most of my family is pro gun and most of them are against anything that would make access more difficult
2	i feel like there are a lot on the right who would agree w background checks to keep their guns though
2	but why
1	I feel like 60-70 but 90 just seems too high
3	because it delays their safely operating the weaponry and decreases their access to something that allows them to hunt
1	im good with false if eveyone else is
5	i think the statements true
3	they know that the world is going to shit and claim, kinda rightly so, that it would just make it more difficult for those who are safe with guns to be able to have them if there is ever a circumstancer that arises that they would need it
1	I think background checks are good but theirs a part of the country that just doesnt want any restrictions

C.3 Beliefs

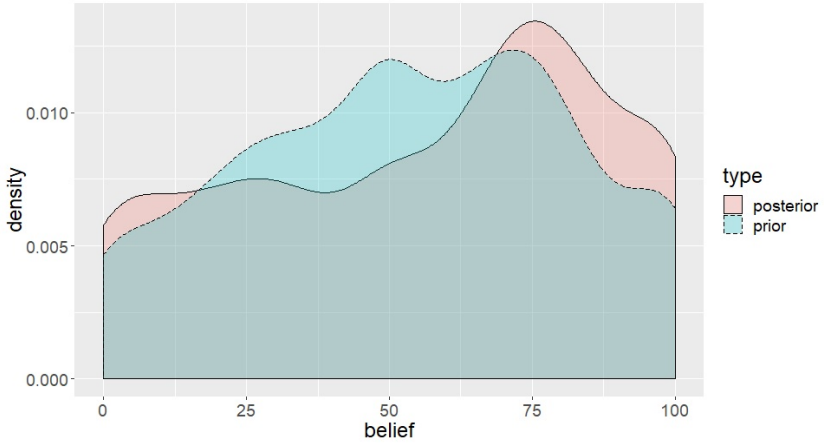
Table C7: Testing for order effects on beliefs

	Dependent variable:			
	Prior		Posterior	
	Set F	Set P	Set F	Set P
P statements first	1.181	−0.046	2.176	0.505
	(1.319)	(1.372)	(1.426)	(1.475)
Constant	55.261	48.910	57.634	47.698
	(0.912)	(0.949)	(0.986)	(1.020)
Observations	1,840	1,840	1,840	1,840
R^2	0.0004	0.00000	0.001	0.0001

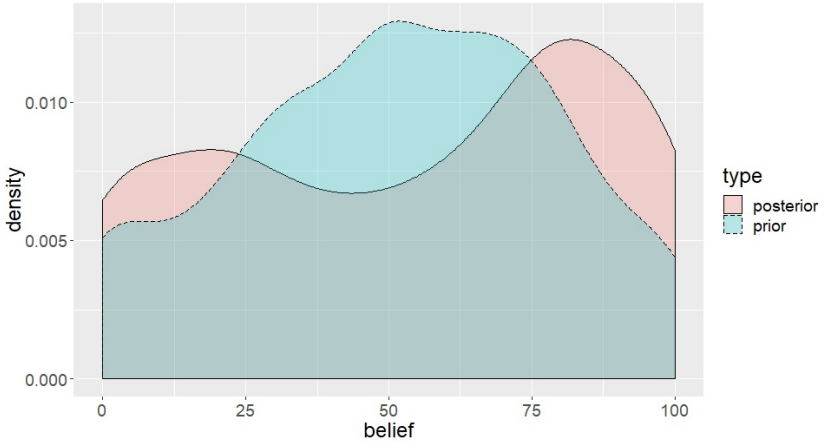
p<0.1; p<0.05; p<0.01

Figure C1: Kernel densities of prior and posterior beliefs for set F statements

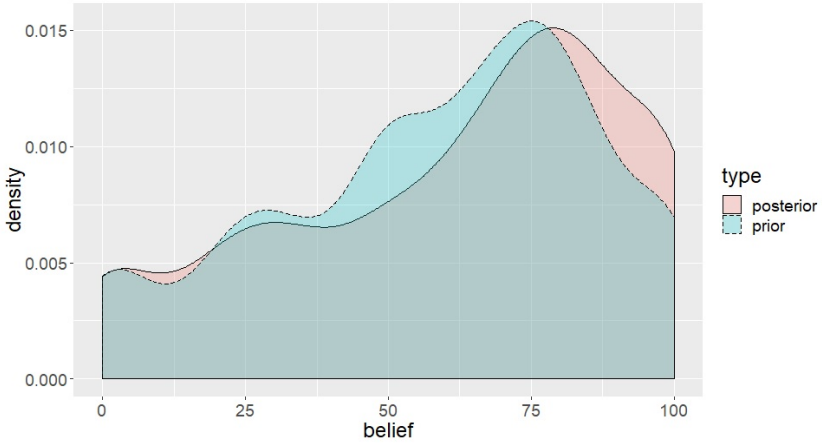
(a) True Set F statements in Vote



(b) True Set F statements in Chat



(c) False Set F statements in Vote



(d) False Set F statements in Chat

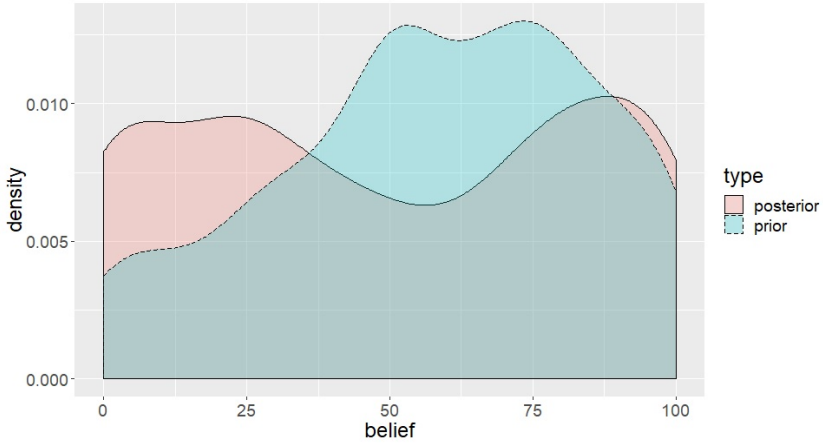
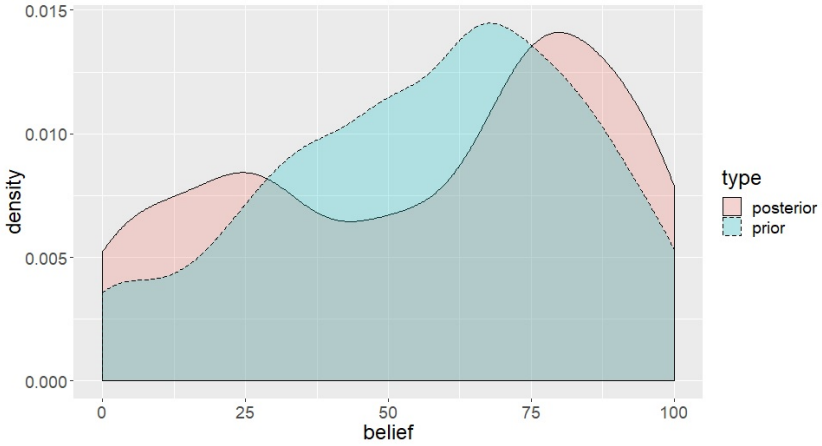
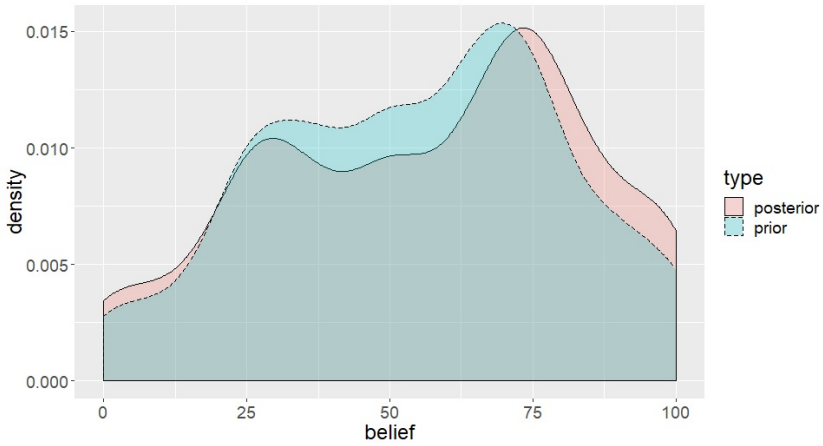


Figure C2: Kernel densities of prior and posterior beliefs for set P statements

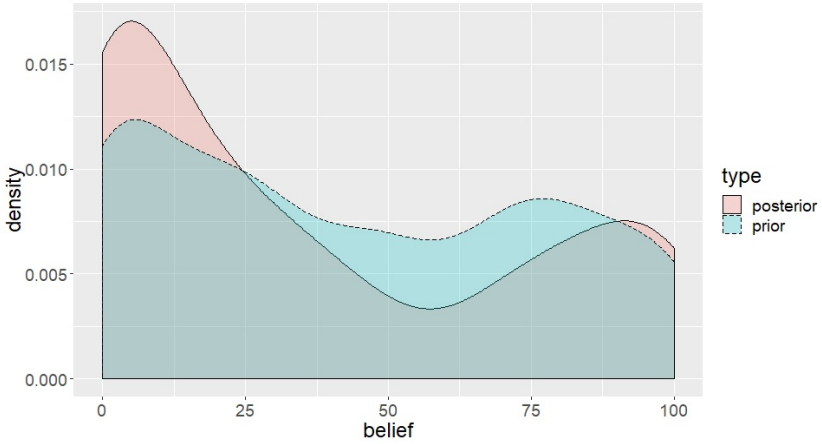
(a) True Set P statements in Chat



(b) True Set P statements in Vote



(c) False Set P statements in Chat



(d) False Set P statements in Vote

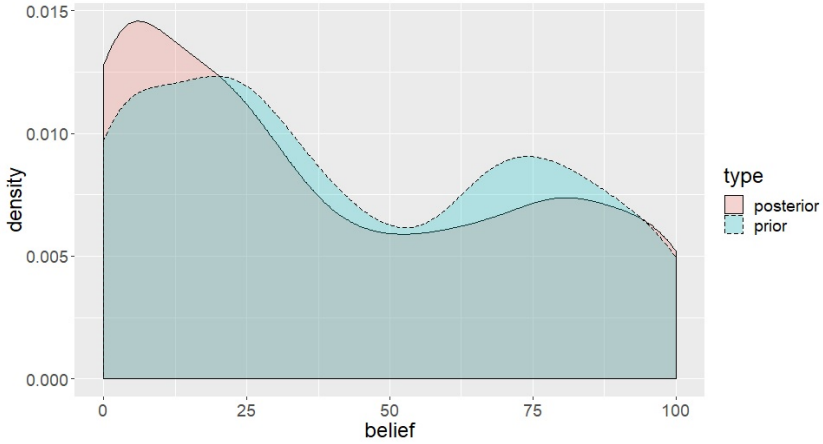


Table C8: Effect of chat depends on item difficulty even after excluding extremely easy and extremely hard items

	Dependent variable:			
	Prior Belief	Posterior Belief	Individual Vote	Group Decision
	(1)	(2)	(3)	(4)
Chat	0.001	0.018	0.059	0.040
	(0.012)	(0.021)	(0.031)	(0.045)
Chat × Easy	−0.002	−0.035	−0.035	−0.171
	(0.020)	(0.021)	(0.034)	(0.072)
Chat × Hard	−0.001	0.061	0.038	0.175
	(0.018)	(0.020)	(0.032)	(0.067)
Easy	0.124	0.168	0.224	0.379
	(0.020)	(0.037)	(0.048)	(0.071)
Hard	−0.105	−0.154	−0.152	−0.261
	(0.019)	(0.034)	(0.045)	(0.066)
Constant	0.516	0.537	0.516	0.522
	(0.012)	(0.026)	(0.034)	(0.044)
Observations	5,110	5,110	5,110	1,022
Log Likelihood	−590.947	−948.149	−3,444.318	−640.483

Linear mixed effects model with subject, item, and session random effects in columns 1-3, and item and session random effects in column 4. Items with extremely easy (> 0.8 accuracy) and extremely hard (< 0.2 accuracy) difficulty are excluded.