European Journal of Political Economy 90 (2025) 102768

Contents lists available at ScienceDirect

European Journal of Political Economy

journal homepage: www.elsevier.com/locate/ejpe

Full length article

Getting it right: Communication, voting, and collective truth finding ☆

Valeria Burdea [a],*, Jonathan Woon [b]

[a] LMU Munich, Faculty of Economics, Ludwigstraße 28, 80539, Munich, Germany

[b] University of Pittsburgh, Department of Political Science, Department of Economics (secondary), and Pittsburgh Experimental Economics Laboratory, 4437 Wesley W. Posvar Hall, Pittsburgh, PA 15260, USA

ARTICLE INFO

Dataset link: https://osf.io/kbe9n/

JEL classification:
D70
D72
D83

Keywords:
Collective decisions
Voting
Communication

ABSTRACT

We conduct an experiment to examine how communication affects the accuracy of collective judgments in small groups evaluating the truth of politically relevant facts and statements. We find that communication improves the accuracy of the group when individuals are more likely to be wrong, but reduces it when individuals are more likely to be correct—a pattern that reveals how deliberation can both clarify and confuse. Communication influences not only group accuracy but also individual belief updating. When groups vote without prior discussion, individuals appear to interpret others' votes as mildly informative signals and update their beliefs about the likelihood that the statement is true accordingly. However, when communication occurs before voting, this pattern disappears, suggesting that social cues conveyed through discussion override the informational value of votes. Our analysis of chat transcripts reveals that group members use communication to share factual knowledge and engage in interactive reasoning, especially for difficult items. These findings highlight that while deliberation can facilitate truth-seeking, it can also undermine accuracy when consensus builds around mistaken beliefs.

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. Collective judgments can take different forms, such as binary (e.g. true/false), categorical (e.g. selecting the most plausible explanation) or numerical (e.g. estimating quantities), and the nature of the task influences how groups aggregate information. Our study focuses specifically on binary truth-evaluation tasks, enabling a clear test of how communication influences accuracy when groups must decide whether factual statements are true or false.

☆ We thank participants at the 2019 Annual Meeting of the American Political Science Association, Washington, D.C and seminar participants at Florida State University and Warwick Business School. 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).

* Corresponding author.
E-mail addresses: valeria.burdea@econ.lmu.de (V. Burdea), woon@pitt.edu (J. Woon).

https://doi.org/10.1016/j.ejpoleco.2025.102768
Received 31 December 2024; Received in revised form 27 September 2025; Accepted 3 October 2025
Available online 15 October 2025

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 et al., 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, 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 et al., 2003; Lorenz et al., 2011, 2015; Jasny et al., 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 compared to voting alone? We designed and conducted an incentivized lab experiment to investigate the effect of communication on the accuracy of small-group majority opinions regarding a variety of real-world, politically-relevant facts. 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). Moreover, individual 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). Our research design enables us to better understand the extent of these issues in small group decision making processes.

Participants in our experiment were presented with true or false statements, half of which were related to political or general knowledge, while the other half involved claims made by real politicians. This creates an environment with a wide range of factual statements, increasing the external validity of our study. In all treatments we first elicited individuals' prior beliefs about the likelihood that each statement was true. This was followed by the group decision stage where participants voted for each statement being true or false and the group decision was decided by majority rule. Across treatments, we manipulated whether this stage was preceded by an unstructured, free-form chat with the members of one's group. Finally, we presented participants with each statement again and information about how the other group members voted, and elicited participants' posterior beliefs about the likelihood a statement is true. This design allows us to test not only the effect of communication on group decisions, but also the dynamics of individual beliefs across treatments.

We find that communication improves the accuracy of group decisions, but the effect is not uniform. Our rich decision environment allows us to differentiate between hard decision problems where individuals are more likely to be incorrect and easy problems where individuals are more likely to be correct. Compared to a setting where group decisions are taken solely by independent voting, communication increases group accuracy for hard problems, but diminishes it for easy ones.

With respect to individuals' beliefs, we find that these change in response to the information brought by other group members through communication. However, posterior beliefs are more accurate with communication than without only for hard items. To understand how individuals use the available social information, we estimate an empirical belief-updating model and show that, in absence of communication, individuals perceive other people's votes as mildly informative signals, relying mostly on their own priors to form their posterior beliefs. However, the weight put on social information increases when communication is allowed. As suggested by the analysis of the chat transcripts, this can be due to communication being used by subjects not only to simply announce their voting intentions, but also to communicate their knowledge of related facts, raise doubts, and attempt to reason together towards the truth.

Our research contributes to the rich literature on collective decision-making in several ways. First, the findings refine our understanding of how communication affects collective intelligence, showing that its impact depends on the context. Specifically, we demonstrate that task complexity plays a crucial role in shaping the effect of group deliberation, influencing group performance not only through individuals' prior biases but also by shaping the structure of the deliberation process. Second, our research integrates individual and group-level dynamics allowing us to better identify the channels through which group reasoning operates, while changes in posterior beliefs indicate the potential longer-term effects for collective judgments. Third, the analysis of the chat transcripts sheds light on how different modes of communication facilitate collective truth-seeking. Overall, our findings have practical implications for designing decision-making processes in teams and organizations. For example, the results suggest that communication can be particularly valuable when tackling complex problems but should be carefully managed to avoid unnecessary complications for simpler tasks.

Several studies from across the social sciences have been interested in the effects of social information – often obtained through group discussion – on the wisdom of the crowd. Mercier and Claidière (2022) provides a short review of the literature. The empirical

[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 et al., 2019).

findings in this literature mostly point to a positive effect of discussion on individual performance and group outcomes. Studies that find negative effects tend to be theoretical, simulation-based or involve social information that is not obtained through free-form discussion. Moreover, most of these studies examine either numerical estimation tasks or abstract decision-making scenarios, often focusing on large groups (e.g., Mercier and Claidière, 2022) or controlled communication constraints. In contrast, our study examines the impact of direct, unstructured communication in small groups tasked with evaluating politically relevant and general knowledge statements. By incorporating a rich decision environment and emphasizing belief updating dynamics, our work extends these findings to scenarios that mirror real-world group decision-making, where factual uncertainty often intersects with subjective biases.

The most relevant studies to our work are those by Lorenz et al. (2015), Morton et al. (2019), and Goeree and Yariv (2011) which are directly interested in the quality of information aggregation in small groups across settings that vary in the type of social information available. Lorenz et al. (2015) examined variations in decision rules (individual, majority, unanimity), group size, and modes of communication (qualitative or quantitative). However, unlike our study which compares group decisions with and without communication while holding other factors constant, Lorenz et al. (2015) did not isolate the effect of communication itself as some form of communication was possible in all of their treatments. They found that group judgments were worse under majority rule. We show that majority rule can still outperform a no-communication scenario, but only for difficult problems. Interestingly, the under-performance of majority rule in Lorenz et al. (2015) contrasts with (Goeree and Yariv, 2011), where a simple majority rule was at least as efficient as unanimity, both with and without communication. A key distinction between these studies lies in the nature of the task: while (Lorenz et al., 2015) used general knowledge questions with a broad (0–100) answer range, Goeree and Yariv (2011) employed a binary prediction task involving abstract "balls and urns". Our study, therefore aligns more closely with Goeree and Yariv (2011) because our main outcome is also binary (whether a statement is true or false), and they also varied the possibility of group communication. However, a significant difference is the scope of communication. Our context-rich environment enables us to disentangle two distinct channels through which communication affects collective decisions: interactive reasoning and information exchange. As explained in the theoretical framework section, these channels can produce effects that differ markedly from a no-communication protocol.

Morton et al. (2019) investigate the aggregation of information through majority voting when voters are, ex-ante, more likely to be incorrect (i.e. biased) and explore the impact of exogenously provided social information on group decision success. To do so, they used a lab experiment involving a set of general knowledge questions with one correct answer participants had to find from a set of two given options. They showed that, as expected, majority voting lead to better outcomes than individual decisions when voters were unbiased (the "bright side of the vote"), but amplified errors when biases were prevalent (the "dark side of the vote"). The study also highlights the differential effects on group accuracy of two types of social information: the opinion distribution of members of similar groups and the likelihood of success of other similar groups. They found that information on the distribution of opinions amplifies the effects of majority voting, but information on prior success has no effect. While Morton et al. (2019) focus on externally provided social information and its role in sequential voting scenarios, our project extends this line of inquiry by examining the impact of direct, within-group communication on information aggregation and decision accuracy in a rich decision environment. By allowing unstructured group discussions, our study captures the dynamics of collaborative reasoning and belief updating, shedding light on how groups collectively navigate both factual uncertainty and cognitive biases. We show that one way in which the dark side of the vote can be avoided is through pre-vote free-form communication.

The rest of the paper is organized as follows. In Section 2 we present our conceptual framework used to derive hypotheses about the effect of communication. Section 3 details our experimental design and implementation procedures. This is followed by the presentation of the main and exploratory results in Section 4. Section 5 discusses our findings and concludes.

2. Conceptual 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, the logic of the Condorcet Jury (CJ) 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 ($p_i < \frac{1}{2}$), 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 Fig. 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 – e.g., they simply announce their votes (with the same probability of accuracy p_i), but no one changes their opinion – we would expect group decisions reached with and without communication to have the same degree of accuracy. If individuals announced their vote intentions and then voted with the majority opinion, 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 Fig. 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

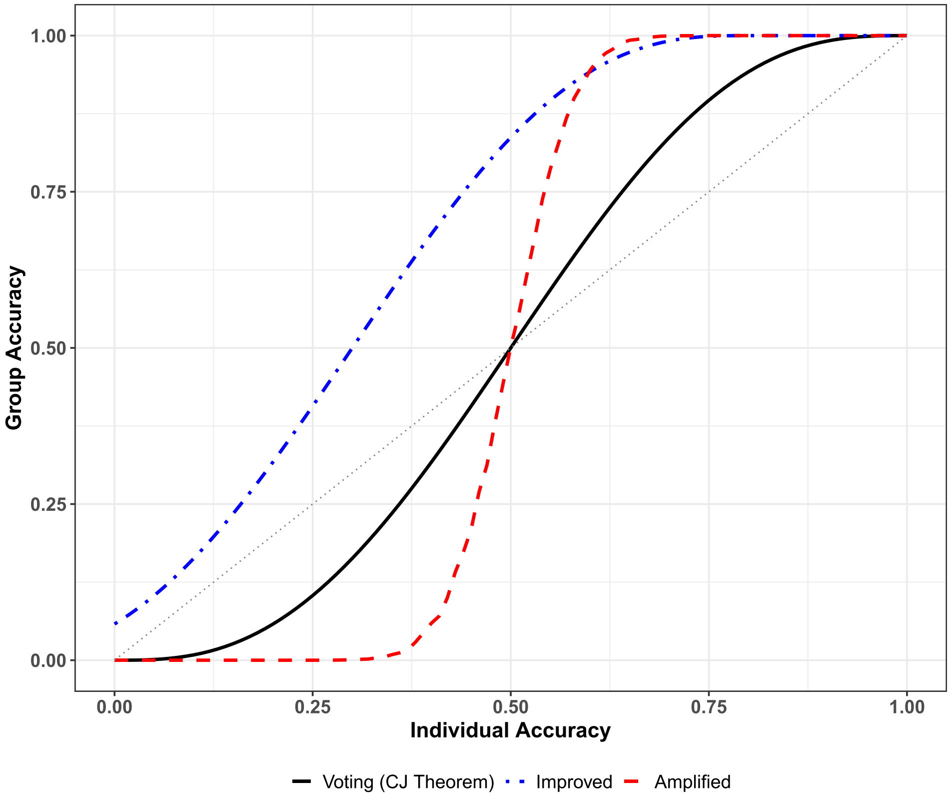

Fig. 1. Group versus individual accuracy and the hypothesized effects of communication for groups of n=5. The solid black curve represents the prediction for the Voting setting without communication according to the Condorcet Jury Theorem. The blue dot-dashed curve ("Improved") represents the hypothesized effect of communication where it enhances individual reasoning prior to voting (with a boost of $\delta = 0.2$), yielding consistently higher group accuracy. The red dashed curve ("Amplified") represents the result of 10000 simulations of group decisions where communication increases the correlation among individual votes without improving individual accuracy, by assuming that members vote on a common signal (accurate with probability equal to the individual accuracy) plus small individual-specific noise (here, the standard deviation of the individual-level noise is $\sigma_i = 0.1$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

another member knows only that the animal flies. On its own, each piece of information is not terribly informative, but by sharing 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 deliberation 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 have 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. We can simulate and visualize this effect by applying an accuracy "boost" of size δ to each individual's prior accuracy (achieved through the collective reasoning enabled by the discussion), before applying the binomial logic of the Condorcet Jury Theorem. Votes are still cast independently. This *improvement* hypothesis is illustrated by the blue dashed–dotted line in Fig. 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.

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. We can model this possibility by assuming that group members rely on a shared signal before voting, plus some individual-level noise, which essentially increases the correlation between individual votes. A simulated implementation of this is shown by the dashed red line in Fig. 1 that lies above the black line for $p_i > \frac{1}{2}$ and below it for $p_i < \frac{1}{2}$.

Finally, we note that all of these effects would be starker with larger group sizes and we illustrate how the steepness of the curves depicted in Fig. 1 increases as we increase group size in the Appendix (Figure C1).

3. Experimental design and procedures

The central task in our experiment was for groups to judge whether real-world statements were true or false. We constructed a pool of 32 statements, divided into two sets of 16: F (factual) and P (politicians' claims). Each set contained eight true and eight false statements (see Table A1 in the Appendix for the full list and sources). Statements were pre-tested to ensure variation in individual accuracy while balancing truth and partisan direction.[2]

Set F consisted of 12 policy-relevant facts commonly used in political knowledge studies (Bullock et al., 2015; Hill, 2017; Robbett and Matthews, 2018), such as changes in unemployment, median income, refugee numbers, or gun-related deaths. Following Hill (2017), most were worded so that, if true, they favored one party. For instance, "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. We also included four neutral facts (three historical, one about airports).

Set P included verifiable claims made by politicians (e.g., Hillary Clinton, Nancy Pelosi, 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), we relied on PolitiFact to assess whether the statements were true or false.[3]

We selected both types to create a diverse pool. Our analysis controls for item characteristics to identify general patterns, but we also check for differences between F and P statements in the Results section.

Experimental structure. Each session consisted of multiple parts. Each part contained 16 rounds, with one statement per round.

- Part 1: Elicitation of prior beliefs.
- Part 2: Group decision task, with majority rule.
- Part 3: Elicitation of posterior beliefs.

In Voting sessions, Parts 4–6 repeated the same sequence with the other set of statements (P or F). Thus, Voting sessions covered both F and P sets, while Chat sessions covered only one set due to the additional time needed for communication. 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.

There were four session types (treatments): Voting-F (VF), Voting-P (VP), Chat-F (CF), and Chat-P (CP). Each varied by whether communication was allowed and which statement set came first. 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.

The structure of the 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.

Group decision task. 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.[4] To prevent possible learning effects (e.g., about others' or one's own accuracy), no interim feedback was provided about accuracy, votes, or earnings. Only after the group decision task concluded did participants learn outcomes. When posterior beliefs were elicited, they saw how their group voted, but not whether the decision was accurate. This task generated the primary outcome of interest, which was whether or not groups and individuals made accurate decisions.

[2] See Appendix Table A2 for variation across all statements.

[3] 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.

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

Table 1
Summary of experiment.

Part	Task	Rounds	Session type			
			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

Note: This table outlines the structure of the experimental sessions. Each session consisted of multiple parts, each involving 16 rounds. Tasks included eliciting prior beliefs (Part 1 and Part 4), making group decisions (Part 2 and Part 5), and eliciting posterior beliefs (Part 3 and Part 6). The four session types are abbreviated as follows: VF = Voting treatment, factual statements first; VP = Voting treatment, politician statements first; CF = Chat treatment, factual statements only; CP = Chat treatment, politician statements only. In the Voting treatments (VF and VP), participants completed all six parts, evaluating both sets of statements (F and P) in randomized order. In the Chat treatments (CF and CP), only one type of statement was evaluated in a session.

Communication treatment. In Voting sessions, no communication was allowed. In Chat sessions, groups had up to three minutes to exchange free-form text messages before voting. All members could view the ongoing chat history, but could not contact members of other groups. Instructions restricted identifying information and offensive language, other than that, no instructions were provided about how subjects should use the chat. Groups could exit the chat early by unanimous agreement. 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.

Belief elicitation. We measured probabilistic beliefs using the crossover elicitation method.[5] Subjects reported a number $B \in [0, 100]$ representing the likelihood a statement was true, using a slider (0 = certain false, 100 = certain true, 50 = uncertainty).[6]

Payoffs depended on B, a random number $W \in [1, 100]$, and the actual truth of the statement. In the experiment, participants were told that W represented the number of "winning lottery tickets" in the round and that its value was drawn uniformly from 1 to 100. If $B \geq W$, the subject earned \$1 if the statement was true. If $B < W$, the subject earned \$1 with probability $W/100$. Reporting B truthfully maximized expected payoff, making the method incentive compatible.[7]

We refer to the first elicitation as the prior belief (slider default set to $B = 50$). After the group task, we elicited posterior beliefs, with the slider default set to the subject's own prior. At this stage, subjects also saw their group's vote (and in Chat sessions, their group's discussion). The change in beliefs before and after group members were exposed to social information represents one of our secondary outcome of interest.

Our experiment is therefore designed to test the effect of group communication on the quality of group and individual decisions. The richness of our decision environment, with a wide variety of statements that differ in content and difficulty, also allows us to examine how the effect of communication varies across items that differ in individuals' prior accuracy. This natural variation enables us to explore a key dimension identified in our conceptual framework: the role of prior individual accuracy in shaping the effects of communication on group decisions. As illustrated in Fig. 1, whether communication improves or harms group performance may depend on how accurate individuals are ex-ante. Thus, by leveraging variation in belief-based accuracy across items, we can test whether the hypothesized amplification or improvement mechanisms operate differently when individuals are more or less likely to be correct. We can thus provide exploratory insights into when communication is most beneficial or potentially counterproductive.

[5] The crossover method is a stochastic version of the BDM mechanism (Savage, 1971) 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). We chose this 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, using only simple ordinal comparisons.

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

[7] 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). 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 compatibility 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).

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	1035	630	1380	825	1265	705	3680	2160
N groups	207	126	276	165	253	141	736	432

Implementation. 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.[8] Each session lasted approximately 90 min and average payments amounted to $16.

4. Results

We divide the data analysis into two categories. The primary analysis focuses on our main outcome variable–individual and group decision accuracy. The secondary analysis examines additional outcomes, namely participants' prior and posterior beliefs and the content of communication. This secondary analysis helps us explore potential mechanisms underlying the primary results. Within the primary analysis, we further distinguish between main and exploratory components. The main analysis addresses the overall effect of our manipulated variable – communication – on decision accuracy. The exploratory analysis investigates heterogeneity in the data, particularly variation in statement difficulty as implied by participants' priors. We classify this as exploratory because, while we anticipated heterogeneity along this dimension, we could not determine in advance which statements would fall into which category, given our participants' prior knowledge.

4.1. Accuracy

Main analysis. 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.

Exploratory analysis. 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 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.[9] 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.

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

[8] See Appendix for a full list of sample demographics.

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

Table 3
Effect of chat depends on item difficulty.

	Dependent variable:			
	Prior belief (1)	Posterior belief (2)	Individual vote (3)	Group decision (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	5840	5840	5840	1168
Log Likelihood	−612.026	−961.115	−3769.208	−675.363

Note: The table presents the results from linear probability models with subject, item, and session random effects in columns 1–3, and item and session random effects in column 4. *p<0.1; **p<0.05; ***p<0.01.

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 linear probability 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. This approach allows us to 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).[10] The results of this analysis are presented in Table 3 and provide rigorous support for the basic conclusions drawn from the disaggregated data.

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 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 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 C1 in Appendix for the corresponding regression analysis). The effect is also robust to classifying items as easy, medium or hard, for each group independently, using the group members' belief-based accuracy measure (see Table C2 in Appendix). The effect is also directionally unaffected by separating between the type of statement (P or F).[11]

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.*

[10] In particular, we estimate the following linear mixed effects model:

$$\Pr(y_{ijs} = 1) = \beta_0 + \beta_1 \text{Chat}_i + \beta_2 \text{Easy}_j + \beta_3 \text{Hard}_j + \beta_4 (\text{Chat}_i \times \text{Easy}_j) + \beta_5 (\text{Chat}_i \times \text{Hard}_j) + u_j + v_s + \varepsilon_{ijs}$$

where y_{ijs} is a binary indicator for whether the individual (or group) i answered item j correctly in session s, Chat_i indicates whether the subject was in the communication treatment, Easy_j and Hard_j are item difficulty indicators (with medium as the baseline), and the interaction terms allow for treatment effects to vary by difficulty. The model includes random intercepts $u_j \sim \mathcal{N}(0, \sigma_u^2)$ for items and $v_s \sim \mathcal{N}(0, \sigma_v^2)$ for sessions. For individual-level outcomes, we also include subject-level random intercepts. The error term $\varepsilon_{ijs} \sim \mathcal{N}(0, \sigma^2)$ captures idiosyncratic variation.

[11] When we do so, however, 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 C5 and C6 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 C3 and C4 in the Appendix).

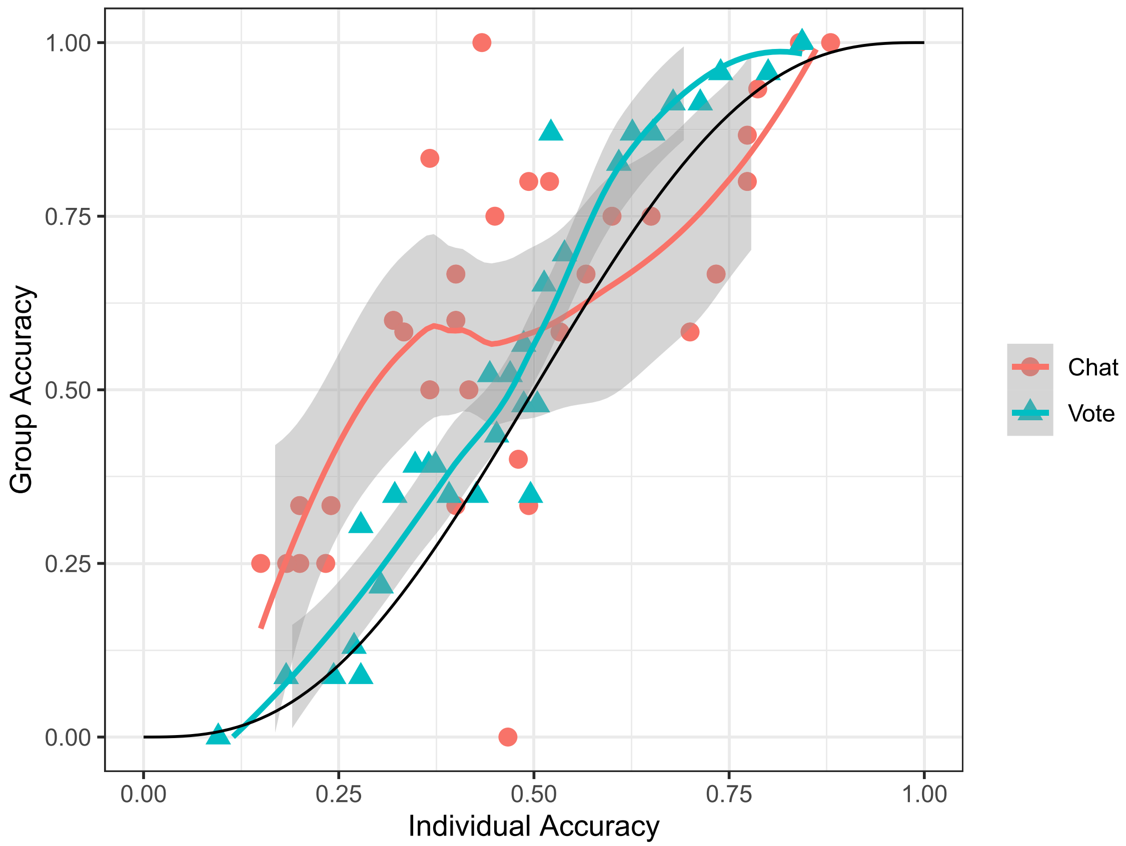

Fig. 2. Group versus individual accuracy.

We generate the empirical analogue of Fig. 1 by disaggregating the data further, to the item level, and plotting group accuracy against individual (belief-based) accuracy. The results are shown in Fig. 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] 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 Fig. 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 the 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 Fig. 1), we did not expect communication to decrease performance.

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

[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.

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 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]

To better understand how group members integrate and respond to social cues during decision-making, and whether the driver of the differential effect of communication on group decisions is more likely simple information exchange, or a more complex form of interactive reasoning, we first turn to an analysis of the belief updating process.

4.2. Beliefs

Beyond influencing group decisions, communication may also shape how individuals revise their beliefs. To investigate how social information affects individual learning, we analyze belief updating in the presence and absence of communication. We ask whether individuals treat others' votes as informative signals, and how this changes when communication occurs before voting. These comparisons allow us to estimate how informative individuals perceive others' opinions to be, and whether communication conveys more than what votes alone reveal. This analysis helps identify the mechanisms through which communication shapes judgment, clarifying the informational and interpretive dynamics that later unfold in open-ended group discussions.

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 kth 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 kth 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

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

The corresponding empirical specification is

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

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 Eq. (2) separately for the Vote and Chat treatments, with random effects for sessions, and statements included in the model.[14]

[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/).

[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.

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

	Dependent variable:			
	logit(posterior)			
	(Vote)	(Chat)	(Vote)	(Chat)
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	3680	2053	3680	2160
Log Likelihood	−6013.898	−3668.233	−6271.832	−4349.438

Note: 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. *p<0.1; **p<0.05; ***p<0.01.

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 Bayesian framework. This is appropriate in the Vote 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.

Result 2. *Participants in the Vote treatment take into account their group members' prior beliefs in their belief updating process, but they do so conservatively, relying primarily on their priors.*

In the Chat treatment, the observed votes are not simply indicators of the independent signals (beliefs) of other group members but instead affected by communication (and therefore correlated). Indeed, Fig. 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] 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 of Table 4 suggest that belief updating

[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 experimental 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.

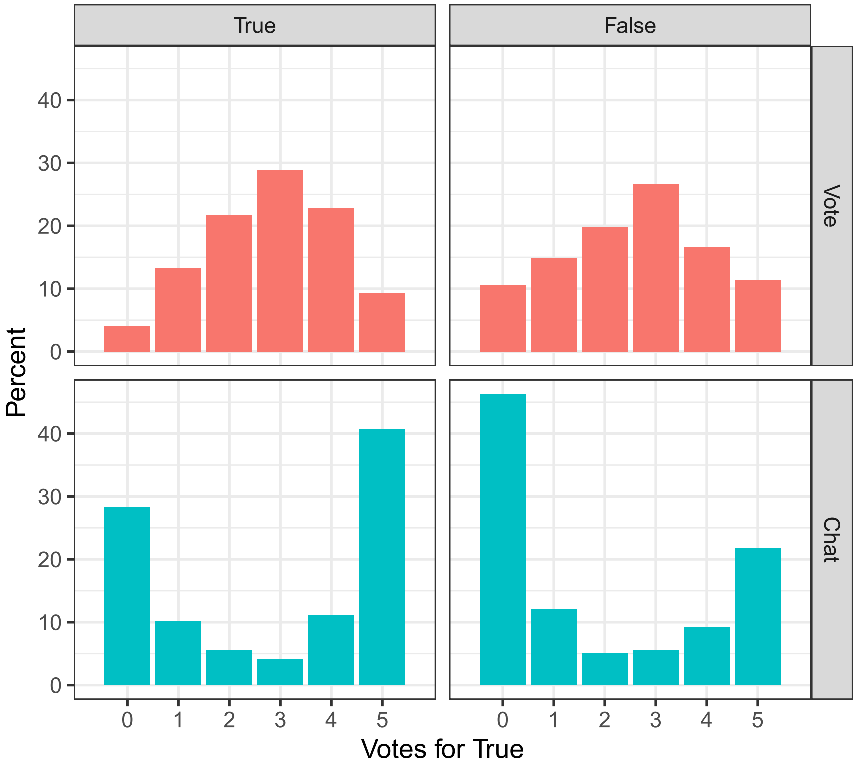

Fig. 3. Communication induces unanimous decisions.

in the Chat treatment differs in two ways from the Vote treatment. First, subjects in Chat compared to Vote rely less on their prior beliefs (which have a smaller coefficient) and more on the available social information. Second, messages in favor of True seem to have a significantly positive effect on people's beliefs while those in favor of False do not.

Result 3. *When updating their beliefs, participants in Chat rely less on their own priors and incorporate the social information more than those in Vote.*

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. However, neither belief (vote proxy) 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 4. *The social information used by participants in Chat when updating their beliefs is different than simply their group members' prior beliefs.*

These findings indicate that communication not only increases the weight participants place on social information but also changes the nature of that information. Rather than functioning as a channel for transmitting prior beliefs, communication alters the informational context through interactive dialogue, shaping how individuals interpret others' views and form their own judgments. To better understand the informational dynamics of communication, we analyze how the structure and substance of group discussions vary with item difficulty and influence collective accuracy.

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

Table 6
Bag-of-words analysis of chat activity and substance.

	Dependent variable:			
	#Messages (1)	#Words (2)	Words/Msg. (3)	%TF Msgs. (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	−1497.838	−2281.206	−842.489	64.615

Note: The table presents the results of linear mixed effects regression models with item and session random effects. Each observation corresponds to one "chat" (a single group discussion of one statement). 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. "#Messages" and "#Words" represent the number of messages and number of words present in a chat. "Words/Msg." represents the number of words per message in a chat. "%TF Msgs." represents the percentage of messages in each chat that include the word "true" or "false" (or some variant such as "t", "f", "tru", "fake", "fals"). *p<0.1; **p<0.05; ***p<0.01.

4.3. What groups say – communication content and collective accuracy

Having shown that communication alters how individuals respond to social information, we now examine the content of group discussions to better understand the mechanisms driving its effects. In particular, we explore how interactive reasoning – through the exchange of perspectives, expressions of uncertainty, and collective interpretation – contributes to group performance, and how these dynamics help explain the heterogeneous effects of communication across varying levels of item difficulty.

Group chats vary widely in their substance. Some consist exclusively of simple declarations of belief, as in the very brief (yet also complete) transcript shown in Table 5.[18]

Other chats are better characterized by more substantive information exchange and an interactive reasoning communication mode. These exhibit richer, more deliberative exchanges, such as one discussing the topic "Hate speech is not protected by the first amendment"[19] which included factual claims (e.g., "the first amendment protects freedom of speech"), expressions of uncertainty (e.g. "Idk if hate speech is free speech"), and arguments (e.g. "that is targetting of people. free speech is more crtical of government and authority wont land you in trouble").[20]

To understand how common it is for groups to engage in processes of interactive reasoning and how this interacts with the difficulty of the task at hand, we use both a quantitative textual analysis based on the frequency of certain communication measures (bag-of-words approach) and a qualitative analysis based on the ratings of two human coders. Table 6 presents mixed effects models for several measures of activity based on the bag-of-words analysis.

We find that the content of communication is closely tied to item difficulty. Harder items elicit longer, more substantive discussions with greater message volume and longer average message length. As a crude measure of the frequency of simple declarative statements of private opinions, 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 6, 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-difficulty items and 7 percentage points for hard items.

[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/.

[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] See Appendix for the full chat on this topic and further examples.

Table 7
Human-coded chat substance.

	Dependent variable:			
	Belief (1)	Fact/Argument (2)	Response (3)	Doubt (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.

Thus, when items are more difficult, there is an increase in chat volume and a corresponding decrease in simple declarations of members' guesses, suggesting a greater likelihood for interactive reasoning. This is further supported when we examine two additional measures of content based on the presence of a handful of keywords that can proxy the presence of (i) simple information exchange (messages containing words such as “think”, “feel” or “believe.”) and (ii) interactive reasoning (messages containing words such as “because”, “so”, “argument”, “but”, “maybe”, “although”, “reason”).[21] Based on these measures, we observe that harder items are associated with an increase in the frequency of interactive reasoning for hard questions (24%) compared to easy questions (16%).

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).[22][23] Table 7 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 6, column 4). The differences in this case is though not significant. Next, as prefaced by the textual analysis, 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). Based on the bag-of-words and human-coded analysis, we can therefore formulate the following result:

Result 5. *The mode of communication is correlated with item difficulty: when items are more difficult, groups exchange more and lengthier messages, including more facts and arguments.*

This pattern provides evidence that groups do not simply share opinions or vote intentions, but adapt their communication style depending on the task. Harder items elicit more substantive exchanges, consistent with interactive reasoning rather than naive signal sharing. This distinction is critical: it suggests that communication enables groups to shift from information aggregation to collaborative inference when individual uncertainty is high. Identifying this shift helps explain why communication improves accuracy for difficult items, and motivates further analysis examining whether specific communicative patterns – such as the presence of high-quality contributions – drive group performance.

We now ask whether these differences in communication help explain why accuracy improves for harder items. In addition to the group-level patterns identified above, we examine the role of individual group members—specifically, whether certain participants exert greater influence on outcomes, and whether their accuracy shapes the direction of that influence. Since communication increases the rate of unanimous group decisions (see Fig. 3), it is possible that some members disproportionately sway the outcome.

[21] In many of the messages, words in category (i) are associated with tentative assertions of factual claims. Examples include: “I think they had a war in like the early 2000 s”, “i feel like we made the transition to oil a while agho”, or “Murders have been decreasing steadily since the 80 s I believe.” Words in category (ii) are suggestive of a group deliberation process. Examples include: “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”. We take into account also variants of these words such as “bc”, “cause”, “though”, etc.

[22] 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.

[23] 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 - $\alpha = 0.544$. 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. 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
Group accuracy, influential members, chat activity, and human-coded chat content.

	Dependent variable: Group accuracy			
	Pooled (1)	Easy (2)	Medium (3)	Hard (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

Note: The table presents the results from linear mixed effects regression models with session random effects. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

If these individuals are well-informed, communication can enhance accuracy; if not, it may amplify error. To explore this possibility, we construct measures of each group's most influential members – based on their chat contributions and prior beliefs – and include these alongside group-level communication metrics in our analysis.

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 confident* member(s) of each group in terms of prior beliefs furthest from $\frac{1}{2}$ for the item under consideration (regardless 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 for these equally talkative/confident members. In addition to the accuracy of key members, we also include the number of accurate prior opinions in the group in the analysis.

Table 8 presents the estimates from a linear mixed effects regression with group accuracy as the dependent variable and 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.

We present the results based on 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). The results using the bag-of-words analysis are in the Appendix (Table C10) and are broadly in line with the human-coded analysis. However, the bag-of-words measures for chat content are too coarse and fail to capture the contextual meaning of the exchanged messages.

The main finding is that certain members of the group appear to be influential, driving the group's decision. Specifically, 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 if their prior is accurate. The coefficient for the first one to speak has a negative sign for easy items and a positive sign for medium and hard ones. 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.

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

Regarding the effect of the chat content, 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 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 6. *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 et al. (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.

These findings suggest that collective decisions can benefit from social information at all levels of item difficulty—so long as it is conveyed in a format that facilitates meaningful interpretation and reasoning. An important direction for future research is whether groups can recognize and select the forms of social information best suited to the task at hand.

Our findings also raise practical questions about how difficulty should be assessed in real-world settings where decisions must be made about whether to allow or encourage group deliberation. In our experiment, item difficulty was defined empirically based on individuals' prior beliefs, which provides a reliable and treatment-independent benchmark. However, this approach requires knowing the "ground truth" and measuring belief accuracy ex-ante, which is rarely feasible in applied contexts. Promising directions for future work would be to develop proxies for difficulty, such as subjective difficulty ratings from a pilot population or using expert judgments, and to endogenize procedural decisions about when to engage in deliberation.

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 et al. (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.

CRediT authorship contribution statement

Valeria Burdea: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jonathan Woon:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ejpoleco.2025.102768.

Data availability

Data and analysis materials are available here: https://osf.io/kbe9n/.

References

Allen, Franklin, 1987. Discovering personal probabilities when utility functions are unknown. Manag. Sci. 33 (4), 542–544.
Austen-Smith, David, Banks, Jeffrey S., 1996. Information aggregation, rationality, and the condorcet jury theorem. Am. Political Sci. Rev. 90 (1), 34–45.
Bartels, Larry M., 2002. Beyond the running tally: Partisan bias in political perceptions. Political Behav. 24 (2), 117–150.
Bazazi, Sepideh, von Zimmermann, Jorina, Bahrami, Bahador, Richardson, Daniel, 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. Br. J. Political Sci. 47 (2), 241–262.
Bullock, John G, Gerber, Alan S, Hill, Seth J, Huber, Gregory A, 2015. Partisan bias in factual beliefs about politics. Q. J. Political Sci. 10, 519–578.
Cooper, David J., Kagel, John H., 2005. Are two heads better than one? Team versus individual play in signaling games. Am. Econ. Rev. 95 (3), 477–509.
Coutts, Alexander, 2019. Good news and bad news are still news: Experimental evidence on belief updating. Exp. Econ. 22 (2), 369–395.
Delli Carpini, Michael X., Keeter, Scott, 1996. What Americans know about Politics and Why it Matters. Yale University Press.
DeMarzo, Peter M., Vayanos, Dimitri, Zwiebel, Jeffrey, 2003. Persuasion bias, social influence, and unidimensional opinions. Q. J. Econ. 118 (3), 909–968.
Fischbacher, Urs, 2007. Z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Gaines, Brian J, Kuklinski, James H, Quirk, Paul J, Peyton, Buddy, Verkuilen, Jay, 2007. Same facts, different interpretations: Partisan motivation and opinion on Iraq. J. Politics 69 (4), 957–974.
Galton, Francis, 1907. Vox populi. Nature 750, 450–451.
Gilens, Martin, 2001. Political ignorance and collective policy preferences. Am. Political Sci. Rev. 95 (2), 379–396.
Goeree, Jacob K., Yariv, Leeat, 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. J. Econ. Behav. Organ. 17 (1), 31–57.
He, Yunwen, Lien, Jaimie W., Zheng, Jie, 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. J. Politics 79 (4), 1403–1418.
Holt, Charles A., Smith, Angela M., 2016. Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. Am. Econ. J.: Microeconomics 8 (1), 110–139.
Janis, Irving Lester, 1982. Groupthink: Psychological Studies of Policy Decisions and Fiascoes, vol. 349, Houghton Mifflin Boston.
Jasny, Lorien, Waggle, Joseph, Fisher, Dana R., 2015. An empirical examination of echo chambers in US climate policy networks. Nat. Clim. Chang. 5 (8), 782.
Jerit, Jennifer, Barabas, Jason, 2012. Partisan Perceptual Bias and the Information Environment. J. 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. J. Educ. Psychol. 16 (5), 300.
Kugler, Tamar, Kausel, Edgar E., Kocher, Martin G., 2012. Are groups more rational than individuals? A review of interactive decision making in groups. Wiley Interdiscip. Rev.: Cogn. Sci. 3 (4), 471–482.
Kuklinski, J.H., Quirk, P.J., Jerit, J., Schwieder, D., 2000. Misinformation and the currency of democratic citizenship. J. Politics.
Landemore, Hélène, 2017. Democratic reason: Politics, Collective Intelligence, and the Rule of the Many. Princeton University Press.
List, Christian, Goodin, Robert E., 2001. Epistemic democracy: Generalizing the condorcet jury theorem. J. Political Philos. 9 (3), 277–306.
Lorenz, Jan, Rauhut, Heiko, Kittel, Bernhard, 2015. Majoritarian democracy undermines truth-finding in deliberative committees. Res. Politics 2 (2), 2053168015582287.
Lorenz, Jan, Rauhut, Heiko, Schweitzer, Frank, Helbing, Dirk, 2011. How social influence can undermine the wisdom of crowd effect. Proc. Natl. Acad. Sci. 108 (22), 9020–9025.
Lupia, Arthur, McCubbins, Mathew D., 1998. The Democratic Dilemma: Can Citizens Learn What They Need to Know?. Cambridge University Press, Cambridge.
Mercier, Hugo, Claidière, Nicolas, 2022. Does discussion make crowds any wiser? Cognition 222, 104912.
Mercier, Hugo, Landemore, Hélène, 2012. Reasoning is for arguing: Understanding the successes and failures of deliberation. Political Psychol. 33 (2), 243–258.
Mercier, Hugo, Sperber, Dan, 2017. The Enigma of Reason. Harvard University Press.
Mobius, Markus M, Niederle, Muriel, Niehaus, Paul, Rosenblat, Tanya S, 2011. Managing Self-Confidence: Theory and Experimental Evidence. Technical Report, National Bureau of Economic Research.
Morton, Rebecca B., Piovesan, Marco, Tyran, Jean-Robert, 2019. The dark side of the vote: Biased voters, social information, and information aggregation through majority voting. Games Econom. Behav. 113, 461–481.
Navajas, Joaquin, Niella, Tamara, Garbulsky, Gerry, Bahrami, Bahador, Sigman, Mariano, 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. Nat. Hum. Behav. 2 (2), 126.
Novaes Tump, Alan, Wolf, Max, Krause, Jens, Kurvers, Ralf HJM, 2018. Individuals fail to reap the collective benefits of diversity because of over-reliance on personal information. J. R. Soc. Interface 15 (142), 20180155.
Robbett, Andrea, Matthews, Peter Hans, 2018. Partisan bias and expressive voting. J. Public Econ. 157, 107–120.
Savage, Leonard J., 1971. Elicitation of personal probabilities and expectations. J. Amer. Statist. Assoc. 66 (336), 783–801.
Stroop, John Ridley, 1932. Is the judgment of the group better than that of the average member of the group? J. Exp. Psychol. 15 (5), 550.
Surowiecki, James, 2004. The Wisdom of Crowds. Anchor.
Woolley, Anita Williams, Chabris, Christopher F, Pentland, Alex, Hashmi, Nada, Malone, Thomas W, 2010. Evidence for a collective intelligence factor in the performance of human groups. Science 330 (6004), 686–688.
Woon, Jonathan, 2019. Political Lie Detection. University of Pittsburgh, Working paper.
Woon, Jonathan, Kanthak, Kristin, 2019. Elections, ability, and candidate honesty. J. Econ. Behav. Organ. 157, 735–753.