Journal of Experimental Social Psychology 96 (2021) 104157

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Case Report

Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion

Ike Silver [a,b,*], Barbara A. Mellers [a,b], Philip E. Tetlock [a,b]

[a] The Wharton School, University of Pennsylvania, United States of America
[b] Department of Psychology, University of Pennsylvania, United States of America

ARTICLE INFO

Editor name: Michael Kraus

Keywords:
Crowd wisdom
Group judgment
Calibration
Teamwork
Confidence
Advice-taking
Estimation

ABSTRACT

'Crowd wisdom' refers to the surprising accuracy that can be attained by averaging judgments from independent individuals. However, independence is unusual; people often discuss and collaborate in groups. When does group interaction improve vs. degrade judgment accuracy relative to averaging the group's initial, independent answers? Two large laboratory studies explored the effects of 969 face-to-face discussions on the judgment accuracy of 211 teams facing a range of numeric estimation problems from geographic distances to historical dates to stock prices. Although participants nearly always expected discussions to make their answers more accurate, the actual effects of group interaction on judgment accuracy were decidedly mixed. Importantly, a novel, group-level measure of *collective confidence calibration* robustly predicted when discussion helped or hurt accuracy relative to the group's initial independent estimates. When groups were collectively calibrated prior to discussion, with more accurate members being more confident in their own judgment and less accurate members less confident, subsequent group interactions were likelier to yield increased accuracy. We argue that collective calibration predicts improvement because groups typically listen to their most confident members. When confidence and knowledge are positively associated across group members, the group's most knowledgeable members are more likely to influence the group's answers.

1. Introduction

Solving problems often requires estimating unknown quantities. How many cases of COVID-19 were transmitted last month? How many square miles of rainforest are left in the Amazon? How many calories did we burn while exercising last night? A surprisingly robust method for tackling such problems is to elicit independent answers from multiple judges and compute a simple average. Across domains, for groups large and small, and even in high-stakes situations, this 'wisdom of crowds' answer frequently outperforms most individual estimates and sometimes even beats the smartest experts (Clemen, 1989; Kurvers et al., 2016; Larrick, Mannes, & Soll, 2011; Surowiecki, 2004).

In reality, however, a central tenet of crowd wisdom – that individual estimates should be independent – frequently breaks down. Groups often want to collaborate and see value in discussion and teamwork. The present research revisits a classic question: Under what conditions should groups work together on estimation problems? Specifically, when does discussion yield judgments that are more accurate than the group's initial average answer? Our proposition is that groups learn more and perform better after discussion when members exhibit stronger *collective confidence calibration* in their initial independent answers. That is, if more knowledgeable members of a group are relatively more confident and less knowledgeable members relatively less confident prior to discussion, subsequent group interaction is likelier to make crowds wiser. In essence, we quantify the impact of 'knowing what you know' on the benefits of group interaction.

1.1. The costs and benefits of teamwork

Although laypeople and practitioners tout the importance of group discussion, scholars often question its value, citing statistical and psychological biases that lead collective judgments astray. From a statistical perspective, interacting groups produce correlated errors that are less likely to cancel out on average. From a psychological perspective, group deliberations often overweight shared information, overlook unique expertise held by knowledgeable individuals (Stasser & Titus, 1985), and neglect the benefits of cognitive diversity (Davis-Stober, Budescu, Dana, & Broomell, 2014; Hong & Page, 2004). Moreover, group

* Corresponding author at: Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104, United States of America.
E-mail address: ISilver@wharton.upenn.edu (I. Silver).

https://doi.org/10.1016/j.jesp.2021.104157
Received 12 November 2020; Received in revised form 6 April 2021; Accepted 22 April 2021
Available online 11 May 2021

interactions often yield processes of herding, bullying, and social signaling (Kerr, MacCoun, & Kramer, 1996; Raafat, Chater, & Frith, 2009) that foster groupthink (Janis, 1982) and bias group judgments in aggregate. Some scholars have even suggested that, relative to aggregated independent answers, face-to-face discussion is antithetical to good group judgment (Armstrong, 2006).

Others believe the benefits of group discussion can outweigh the costs. Indeed, independent estimates do not guarantee unbiased averages. Even individuals working solo fall prey to cognitive biases and rely on misleading heuristics (Prelec, Seung, & McCoy, 2017; Simmons, Nelson, Galak, & Frederick, 2010), and discussion can help surface and correct errors. For example, Minson, Mueller, and Larrick (2017) recently demonstrated that for challenging estimation questions, group discussion can help members identify plausible ranges and rule out farfetched answers that distort group averages. Discussion can also foster engagement and learning (Smith et al., 2009), by encouraging members to articulate and stress-test their rationales (Mellers et al., 2014) and by facilitating the transfer of knowledge from more to less informed group members (Schultze, Mojzisch, & Schulz-Hardt, 2012).

Recently, more elaborate methods have been proposed to capture the benefits of group discussion while avoiding or reducing the costs. These include averaging consensus estimates from smaller subgroup discussions (Navajas, Niella, Garbulsky, Bahrami, & Sigman, 2018), limiting participation in crowd-sourced estimation to those who have demonstrated prior accuracy (Mannes, Soll, & Larrick, 2014), and enlisting third-party observers or unbiased algorithms to facilitate balanced conversation (Dalio, 2017; Regan-Cirincione, 1994). Alternatively, groups can decide whether to discuss or vote independently depending on the task (Laughlin, Bonner, & Miner, 2002; Minson, Meuller & Larrick, 2017). Although such approaches are intriguing in principle, they are challenging to implement. Informal discussion remains by far the most common mode of group problem-solving.

1.2. Present research

What factors predict the effects of group interaction on judgment accuracy? The present research investigates a novel group-level predictor. Specifically, we propose that *collective confidence calibration* predicts the chances that a group's answers to numeric estimation problems will be improved by discussion. Prior literature has defined confidence calibration broadly as the degree of correspondence between subjective confidence and objective accuracy in individuals. That is, calibration is typically treated as individual-level issue of over- or under-confidence (e.g., Alba & Hutchinson, 2000; Moore & Healy, 2008). Here, we suggest that *groups* vary in *collective* calibration.

Collective calibration captures whether individual judgment accuracy and individual confidence are positively or negatively associated among members of a group. See Fig. 1. For example, in one study, we ask participants to independently estimate the distance from Los Angeles to Honolulu and to rate confidence in their estimate, prior to group interaction. We then calculate *collective calibration,* for each group, as the rank-order correlation between initial accuracy (distance between estimate and correct answer) and initial confidence among group members. Our central prediction is that when the association between accuracy and confidence is positive – when more accurate members are relatively more confident in their own judgment and less accurate members are less confident prior to discussion – groups are likelier to benefit from working together. By contrast when confidence and knowledge are *negatively* correlated, groups may listen to their less informed members and discussion may go off the rails (see also, Einhorn, Hogarth, & Klempner, 1977).

Why? Our prediction starts from the premise that confidence tends to be persuasive in discussion: Groups typically listen to their more confident members. Indeed, in laboratory settings, those who behave more confidently are often judged to be more knowledgeable and exert greater influence on group decisions (Price & Stone, 2004; Sah, Moore, & MacCoun, 2013; Zarnoth & Sniezek, 1997). By contrast, timid individuals often have less influence. Although this tendency to herd towards confident people is commonplace, it doesn't necessarily lead groups to better answers. Listening to confidence will benefit groups, we argue, when they are collectively calibrated: when a group's more knowledgeable members – perhaps those who possess domain-relevant expertise or superior logic – also happen to be its most confident members.

In order for groups to arrive at better answers after discussion, members need to identify which people they should believe and follow. Because groups typically *assume* their most confident members to be their most knowledgeable, whether or not confidence and knowledge are *actually* aligned should predict whether answers get better or worse after discussing a problem. We investigate this hypothesis in the context of numeric estimation problems. Results indicate that collective calibration, calculated from the independent estimates and confidence ratings of group members prior to interacting, strongly predicts whether subsequent discussion will improve or degrade judgment accuracy.

Our inquiry contributes to an emerging line of work investigating the effects of metacognition on crowd-wisdom problems (e.g., Hertwig, 2012). For example, recent research has found that accuracy increases when those who feel less knowledgeable can opt out of crowd-sourced tasks (Bennett, Benjamin, Mistry, & Steyvers, 2018), and that expert meta-knowledge can be used to identify common misconceptions held by the crowd (Prelec et al., 2017). However, no previous work has tested whether confidence calibration predisposes groups to work well together. Evidence for this hypothesis would reveal calibration to be not only a hallmark of good individual judgment, but also an ingredient for successful teamwork.

1.3. Overview

We conducted two large laboratory studies. Participants made initial independent estimates for a range of numeric estimation problems, discussed their answers face-to-face in small groups, and then revised their estimates. By measuring the correlation between pre-discussion confidence ratings and initial accuracy across group members and by comparing the accuracy of group judgment before and after discussion, we tested whether collective confidence calibration predicts the effects of social interactions on accuracy.

Study 1 explored the relationship between pre-discussion calibration and post-discussion improvement in judgment accuracy. Study 2 replicated and extended Study 1, testing a possible mediator of this relationship: the capacity of well-calibrated groups to correctly identify their most knowledgeable members. We report all measures, manipulations, and exclusions. Data, code, and appendix materials are available at: https://osf.io/tkgea/?view_only=b576f44d1bd74d2f9a36a248855324d9

2. Method

2.1. Participants

676 participants (66% female, mean age = 24.1, SD = 8.8) were recruited from a business school's behavioral lab. Lab sessions lasted approximately 30 min and were hosted over 12 months. Participants included students, staff, and local community members who could sign-

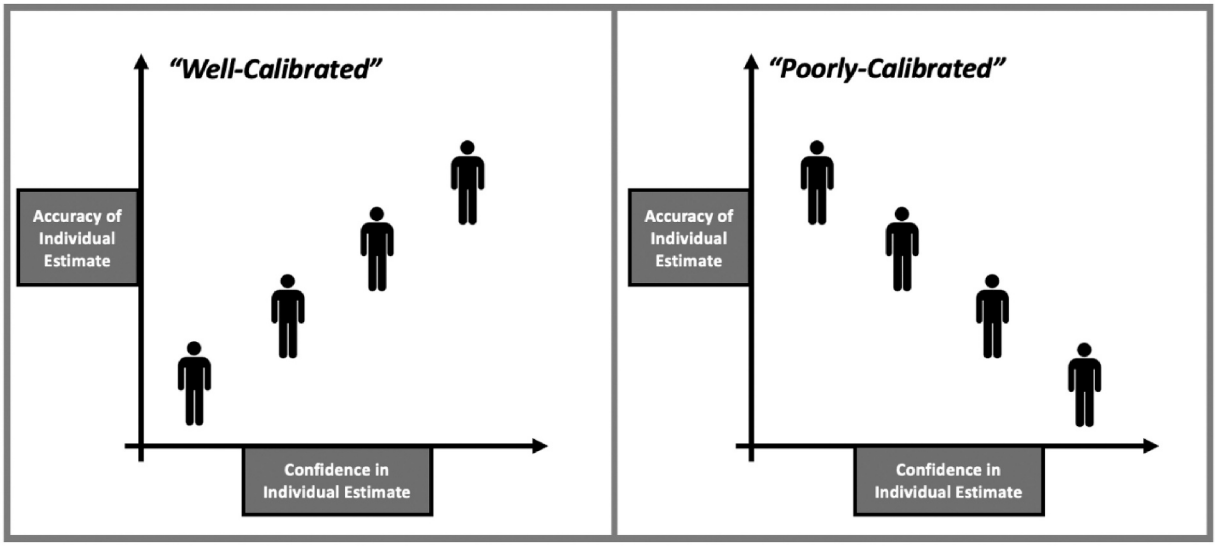

Fig. 1. Collective calibration is a group-level measure of association between accuracy and confidence across group members' answers prior to discussion: Are more confident group members more accurate? Here a perfect positive and perfect negative association are depicted, but the variable can be any value between −1 and + 1, inclusive.

up if they had not participated in a previous session over the prior six weeks. The modal and median number of sessions participants completed was one (min = 1, mean = 1.34, max = 5). Anyone who completed more than one session worked with different teammates and faced different questions.

In Study 1, we requested as many participants as possible to generate a large dataset of groups and estimation questions. In Study 2, we requested 400 additional participants (roughly 80 new groups of 5). Our efforts exhausted the lab's participant pool. The final sample after exclusions[1] included 727 discussions (group-question observations) from 147 teams in Study 1 and 242 discussions from 64 teams in Study 2.[2] Analyses were conducted after data collection was complete.

2.2. Estimation questions

Study 1 used estimations of eight country populations (e.g., Canada, Tanzania) and twelve historical dates (e.g., invention of the printing press). Populations were estimated in millions of residents between 1 and 1000; dates were estimated in years between 1000 and 2018. Study 2 used four new historical-date questions, four questions about direct distances between world cities (e.g., Reykjavik➔Nairobi), and four questions about stock prices for public companies (e.g., Twitter, Chipotle) just prior to data collection (10/24/2018). Distances were estimated in miles between 1 and 10,000; stock prices were estimated in dollars between 1 and 500. In any given lab session, question type and order were constant, but these were varied across sessions.

2.3. Procedures

Participants worked first independently and then in small teams to generate the most accurate answers they could. Cell phones were collected to prevent participants from searching the internet or soliciting outside help. Participants were sorted randomly into groups of 3–6 members and given 2 min to introduce themselves.

Next, they faced a series of estimation questions[3] and followed the same procedure for each. Working first in separate cubicles, participants saw an estimation question and gave their best independent estimate prior to group discussion. They rated their confidence in their own initial, independent estimate and the difficulty of the question on 1 to 7 category-rating scales.[4] Next, they pulled their chairs into a circle to confer in groups. They had three minutes to discuss each question and improve their answer. Groups were not explicitly required to reach consensus (although they often did), and they were not instructed on how to interact. Finally, participants returned to their cubicles, made a second, post-conversation estimate, and rated their confidence in it. Fig. 2 depicts this procedure. After all discussions and estimates were completed, participants indicated whether the group interaction had improved their second estimates on a 5-point scale from −2 "Definitely made my estimations worse" to +2 "Definitely made my estimations better."

Study 2's procedure differed in two notable ways. First, we used more fine-grained measures of participants' perceptions of discussion's impact. Before each discussion, participants were asked, on a 7-point agreement scale, whether they "expect[ed] to get a lot of help from the group discussion to improve [their] estimate." After each discussion, participants indicated how the discussion had influenced their second estimate on a scale from 1 ("Made my second estimate much less accurate") to 7 ("Made my second estimate much more accurate"). We centered these variables around 0, with positive numbers reflecting perceived improvement from discussion. Second, Study 2 measured a hypothesized process variable. After each discussion, participants

[1] *Exclusion Criteria.* We excluded three groups of two members. We also excluded 14 discussions from Study 1 for which all members initially knew the precise correct answer, leaving no room for improvement. These cases were all historical dates (e.g., when the Declaration of Independence was signed). Neither of these exclusions impact the results. We also excluded 39 discussions for which calibration, our primary independent variable, was mathematically undefined, because all group members gave identical confidence ratings.

[2] We recruited additional teams during Study 2 to pilot test training tips to improve/structure conversation. In this report, we only use teams who received no such instructions. However, results are robust to the inclusion of these data. See Appendix.

[3] Typically four, but this varied based on time allotted by the lab.

[4] In early lab sessions, we used several strongly correlated measures of self-confidence. Here we focus on a single-item confidence measure. For details on minor deviations between sessions and associated robustness checks, see Appendix.

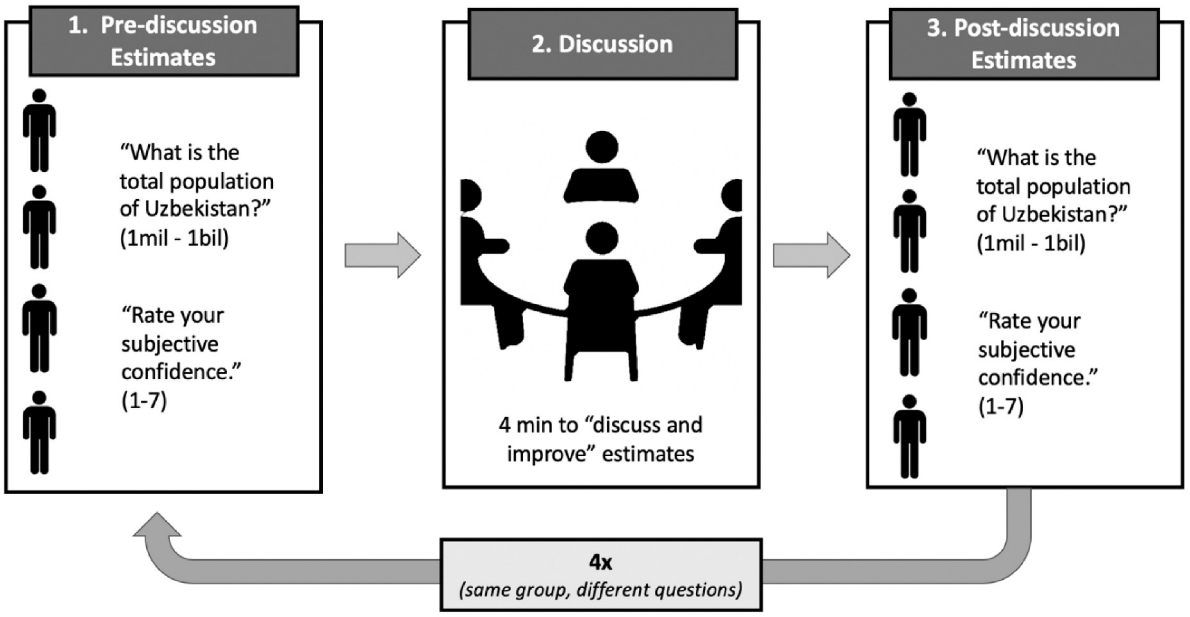

Fig. 2. Stylized overview of research paradigm. We measured group-level collective calibration scores before discussion and used them to predict improvements in accuracy after discussion.

selected the group member who, in their view, had provided the most 'accurate and useful' judgment during discussion. For each team and question (for every discussion), we calculated the proportion of individuals who correctly spotted their most accurate teammate.[5]

2.4. Incentives

Study 1 participants were instructed to make estimates that were as accurate as possible. Study 2 participants were incentivized to give accurate estimates but not to compete with their teammates: A $5 Amazon Gift Card was offered to anyone who provided a post-discussion estimate within circumscribed ranges of the correct answers (e.g., within 5 years for historical dates). We encouraged participants to earn as many bonuses as possible and also help their teammates collect bonuses.

2.5. Measured variables

We operationalized collective confidence calibration – our key pre-discussion predictor – as the rank-order correlation between initial accuracy (absolute distance between estimate and the correct answer) and confidence (1 to 7 confidence rating) across group members' pre-discussion estimates.[6] Because smaller errors mean greater accuracy, we reversed the sign of this variable for clarity, such that positive correlations indicated better calibration – group members with more accurate initial estimations were more confident prior to discussion – and vice versa. Note that for estimations on continuous scales, quantifying *individual* calibration (determining an appropriate 1–7 confidence rating for a certain magnitude of error) is difficult. But we can easily assess whether relatively more confident group members are relatively more accurate to ascertain whether a group is *collectively* calibrated.

We quantified group improvement in two ways. First, we asked whether, for each question, a group's average estimate was closer to the correct answer after discussion than before (1 if yes, 0 if no). This approach treats the 'crowd-wisdom' answer (the average of the group's initial independent estimates without interaction) as a baseline. For convergent evidence, we also examined a secondary continuous outcome variable - the proportion of individuals within a group whose answers improved after discussion. We call these dependent variables "group improvement" and "proportion improved." In all cases, a group or individual was recorded as improved only if the absolute distance between the estimate and the correct answer decreased after discussion. Results were highly similar across measures.[7]

3. Results

We begin with the overarching effects of discussion on accuracy, regardless of calibration, and note participants' overly optimistic expectations about the benefits of discussion. Generally speaking, individuals adjusted their estimates after interacting and consistently reported that group interactions improved their judgment.

In line with classic studies of social influence (e.g., Asch, 1955; Sherif, 1937), participants' estimates were more tightly clustered after interacting, with the group's standard deviation decreasing after 92% and 97% of discussions in Study 1 and Study 2, respectively. Groups also became more confident after discussion, with average confidence ratings increasing after 84% and 90% of discussions in Study 1 and Study 2, respectively. When asked directly, groups also reported that discussion would and did improve their answers. In Study 1, participants reported their perceptions of discussion after all questions and interactions were completed, and we averaged these at the group-level. We found that 146 of 147 groups believed discussion improved their judgment on average.

[5] Participants could not nominate themselves, so we removed the most accurate individual when calculating these proportions.

[6] We do not distinguish here between group-level calibration and resolution, which are sometimes disentangled in studies of individual forecasting and signal detection (e.g., Mellers et al., 2014). Our measure reflects both.

[7] We favor a non-parametric correlation measure of calibration and a binary measure of improvement because (a) they are easily generalized and interpreted across different types of estimation problems with different answer scales and (b) do not rely on assumptions of normality which are frequently violated with small groups and continuous error reduction measures. However, our results are robust to alternative specifications.

In Study 2, groups reported their expectations about discussion's impact before and after each separate discussion. Taking a similar approach, we found that groups expected that discussion would improve their estimates before 75% of discussions, and they said it actually did after 89% of discussions.[8]

In reality, discussion was not nearly as beneficial as participants believed. Only 62% and 55% of Study 1 and Study 2 discussions, respectively, improved group accuracy. Similarly, the average proportion of group members whose individual estimates improved *ex post* was only 47% in Study 1 and 50% in Study 2.

Table 1 displays the effects of group discussion on accuracy. Rows represent estimation questions sorted by rates of improvement. Columns are percentages of groups that improved after discussion, average calibration over groups, and the average correlation between calibration and two measures of improvement over groups (percentage of groups that improved and average proportions of individuals who improved) for each question. Across both studies, discussion increased accuracy more often than chance for 11 questions ($ps < 0.05$), decreased accuracy more often than chance for 6 questions ($ps < 0.05$) and had no detectable effect for the remaining 15 questions.[9]

Our key hypothesis was that better collective confidence calibration prior to discussion would be associated with higher likelihoods of improvement after discussion. We tested this proposition by (a) examining trends over the 32 estimation questions and (b) modeling effects of calibration for all 969 observed discussions.

3.1. Calibration predicts post-discussion improvement – Analysis by question

If better calibrated groups benefit more from discussion, we should see a positive relationship between rates of post-discussion improvement for a given question and average pre-discussion calibration scores of groups facing that question. We found exactly that. Fig. 3 displays the relationship between average calibration (over groups) and the percentage of groups whose average estimate improved, for each of the 32 questions used in Studies 1 and 2 ($r = 0.61$, $p < .001$; Cols 2 & 3, Table 1).

The same positive relationship between pre-discussion calibration and post-discussion improvement also emerged *within* questions. To illustrate, we treated each question as its own weakly powered experiment and computed for each the correlation between pre-discussion calibration and the two post-discussion improvement variables (binary improvement in a group's average answer and the proportion of individuals in a group whose answers improved after discussion). These correlations appear in columns 4 and 5 of Table 1. Positive correlations imply that, for a given question, groups exhibiting higher pre-discussion calibration were likelier to benefit from discussion. Taking an approach akin to meta-analysis, we tested for effects of calibration across these 32 point-estimates. We found a positive correlation on average between pre-discussion calibration and post-discussion group improvement ($M_r = 0.10$, $t(31) = 2.40$, $p = .023$) and a positive correlation on average between pre-discussion calibration and the proportion of individuals who gave more accurate estimates after discussion ($M_r = 0.12$, $t(31) = 3.20$, $p = .003$).

Our proposed mechanism for collective calibration is the enhanced ability of well-calibrated groups to identify their most accurate and knowledgeable members. We measured this variable in Study 2 by asking participants to report, after each discussion, who on their team provided the most accurate input and then calculating the proportion of members who answered correctly. If better calibration is associated with an enhanced ability to identify accuracy, we should find positive correlations between calibration scores and the proportion of group members who recognized the most accurate teammate. This pattern emerged for 9 of 12 estimation problems (Column 6, Table 1). In sum, better calibrated groups were likelier to benefit from group discussion and to recognize their most accurate members both within and across questions. Next, we turn up the microscope and treat each discussion as the unit of observation.

3.2. Calibration predicts post-discussion improvement – analysis by discussion

To test our predictions at the discussion-level, we used mixed-effects regressions to predict the two accuracy-improvement variables from pre-discussion calibration scores, while controlling for effects of question, clustering by group, and group size. We used logistic regressions to predict binary improvement in the group's average answer, and linear regressions to predict the proportion of group members who improved after discussion. Model specifications included random intercepts for groups and questions, and fixed effects for question type (e.g., dates, populations, etc.) and group size.

In Study 1 with 727 discussions, greater collective calibration predicted binary improvement in the group's average ($B = 0.40$, SE = 0.16, *Wald* $Z = 2.51$, $p = .012$) and the proportion of members who improved ($B = 0.046$, SE = 0.017, $t(705.9) = 2.80$, $p = .005$). In Study 2 with 242 discussions, group improvement and proportion of individuals who improved were again associated with greater pre-discussion calibration ($B = 0.49$, SE = 0.25, *Wald* $Z = 1.97$, $p = .049$; and $B = 0.075$, SE = 0.026, $t(221.2) = 2.92$, $p = .004$). Pooling data from Studies 1 and 2 yields similar results,[10] at predictably stronger levels of significance (binary improvement: $B = 0.42$, SE = 0.13, *Wald* $Z = 3.17$, $p = .002$; proportion of individuals who improved: $B = 0.053$, SE = 0.014, $t(928.6) = 3.83$, $p < .001$).

3.3. Robustness checks

We propose that collective calibration predicts whether discussion will yield greater accuracy compared to the groups' initial independent judgments. If true, collective calibration (the *alignment* of accuracy and confidence across group members) should predict *improvement* after taking initial accuracy and confidence into account. Indeed, it does. Similar models that control for the accuracy of the group's initial average answer, the accuracy of the best individual answer, the accuracy of the worst individual answer, the initial spread of individual answers, the group's initial confidence, and group perceptions of question difficulty continue to show significant predictive effects of calibration, with effect size estimates virtually unchanged (for details, see Appendix).

3.4. Psychological mechanism

We argue that pre-discussion calibration should be associated with post-discussion improvement because: (a) better calibrated groups will more reliably identify their most knowledgeable members during discussion; and (b) identification of accurate teammates should increase the likelihood of post-discussion improvement. Study 2 tested this process, with 242 discussions from 64 teams. Using a linear mixed-effects regression model with the same controls as our discussion-level models, we predicted the proportion of group members identifying their most accurate teammate from collective calibration. Better calibrated teams were indeed more successful at identifying their most

[8] *t*-tests comparing groups' average perceptions of discussion to the midpoint of the scale find significant effects in all cases, $ps < 0.001$.

[9] *p*-values represent two-tailed *Z* tests of rates of group improvement by question against chance (50%).

[10] A sensitivity analysis implemented in G*Power indicates that this central test of our theory is powered to detect a minimum effect size of $B = 0.34$ for our primary binary dependent variable. McFadden's R^2 was used to approximate the explanatory power of our additional controls.

Table 1
Summary statistics on average calibration, average improvement, and correlations between collective calibration and group improvement (binary) or proportion improved by question from Studies 1 and 2. Significance codes: ^ $p < .1$, * $p < .05$, ** $p < .01$, *** $p < .001$.

	# of Groups	% Groups Improved	Avg Calibration	Corr. (Calibration, Group Improvement)	Corr. (Calibration, Prop Improved	
STUDY 1 – Questions						
Year Columbus Sets Sail for America	36	97% ***	0.61	−0.09	−0.04	
Year Declaration of Independence Signed	28	93% ***	0.65	0.15	0.06	
Year Printing Press Invented	39	90% ***	0.14	0.28	0.01	
Year DNA Discovered	16	88% **	0.14	−0.32	−0.22	
Year University of Pennsylvania Founded	38	84% ***	0.46	0.29	0.31	
Year First Modern Olympics Held	39	82% ***	0.29	0.33	0.19	
Year Genghis Khan Born	38	76% **	0.26	0.02	0.09	
Population of Canada	42	76% **	0.1	0.14	0.29	
Year First Immunization Delivered	16	75% ^	−0.28	0.03	−0.33	
Year of First Airplane Flight	15	73%	0.07	0.2	0.18	
Population of United States	40	73% **	0.38	0.08	0.42	
Population of Germany	46	65% ^	0.28	0.07	0.18	
Year Beethoven Dies	39	64%	0.13	0.19	0.44	
Population of Australia	46	57%	0.12	0.04	−0.13	
Population of Democratic Republic of Congo	48	56%	0.04	−0.13	−0.14	
Year of Great Earthquake of Kanto	32	31% ^	0.04	0.14	0.26	
Year Flushing Toilet Invented	37	30% *	−0.01	0.01	0.09	
Population of Uzbekistan	42	29% **	0.05	−0.01	0.07	
Population of Tanzania	43	28% **	−0.08	0.02	−0.11	
Population of Madagascar	47	23% ***	−0.06	0.19	0.28	Corr. (Calibration, Prop Identifying Most Accurate)
STUDY 2 – Questions						
Year University of Cambridge Founded	24	88% ***	0.14	−0.25	−0.07	−0.03
Year Ottoman Empire Fell	24	75% *	0.49	0.38	0.39	0.31
Price Per Share Twitter Inc.	21	71% ^	0.04	0.47	0.44	0.1
Distance from Cairo to Sydney	17	71%	0.06	0.19	0.07	0.01
Distance from LA to Honolulu	18	67%	0.2	0.2	0.4	0.2
Price Per Share Home Depot Inc.	20	60%	0.07	−0.37	0.07	−0.25
Year Da Vinci's Last Supper Completed	24	58%	0.27	0	0.38	0.46
Distance from St. Petersburg to Beijing	19	47%	−0.26	0.55	0.36	0.28
Year of Great Lisbon Earthquake	16	44%	0.09	0.33	0.14	0.33
Price Per Share Adobe Inc.	21	33%	−0.01	0.18	0.17	0.04
Distance from Reykjavik to Nairobi	18	17% **	−0.17	0.24	−0.19	0.41
Price per Share Chipotle Mexican Grill Inc.	20	15% **	0.06	−0.44	−0.12	−0.29

accurate member ($B = 0.079$, SE = 0.039, $t(229.7) = 2.044$, $p = .042$). This recognition-of-accuracy variable in turn predicted both improvement variables: group improvement on average ($B = 1.63$, SE = 0.44, Wald Z = 3.70, $p < .001$) and proportion of individuals who improved ($B = 0.22$, SE = 0.042, $t(235.5) = 5.24$, $p < .001$), using similar models.[11]

Finally, we investigated recognition-of-accuracy (measured in Study 2 only) as a possible mediator of the relationship between pre-discussion collective calibration and post-discussion improvement. To do this, we computed average calibration, average proportion of group members recognizing the most accurate person, and proportion of questions for which the group's average answer improved (the mean of our binary improvement variable, i.e., the group's improvement rate) across questions given to each group. The model included the group's improvement rate as the outcome variable, average calibration as the independent variable, and average proportion identifying their most accurate teammate as the mediator. This specification reduces statistical power because it uses only one observation per group (64 total observations), but it also helps us account for repeated measures. Using a bootstrapping procedure with 10,000 samples (Preacher & Hayes, 2004), we detected a significant indirect effect of the mediator – recognition of the most accurate member ($B = 0.073$, SE = 0.05, 95% CI [0.002, 0.18]). We found similar results using the average proportion of individuals who improved as the outcome variable ($B = 0.09$, SE = 0.05, 95% CI [0.013, 0.19]). These analyses are consistent with the hypothesis that pre-discussion calibration predicts post-discussion improvement in part because well-calibrated groups are better equipped to identify accuracy in their ranks.

3.5. General discussion

Working together is ubiquitous, seemingly necessary, and, with the advent of collaboration technology, increasingly convenient. Yet group discussion may not always meet our expectations. Although our participants consistently thought that group discussion would improve accuracy, reality suggested otherwise. Group discussion sometimes

[11] In line with prior work on the influence of confidence in group discussion (e.g., Zarnoth & Sniezek, 1997), we find that when asked to pick out the most accurate person in their group, participants were about as likely to identify the most *confident* peer as they were to identify the most *accurate* peer (on average, 44% of the group picked out the most confident vs. 43% identifying the most accurate). In our context, identifying confidence alone did not predict group improvement.

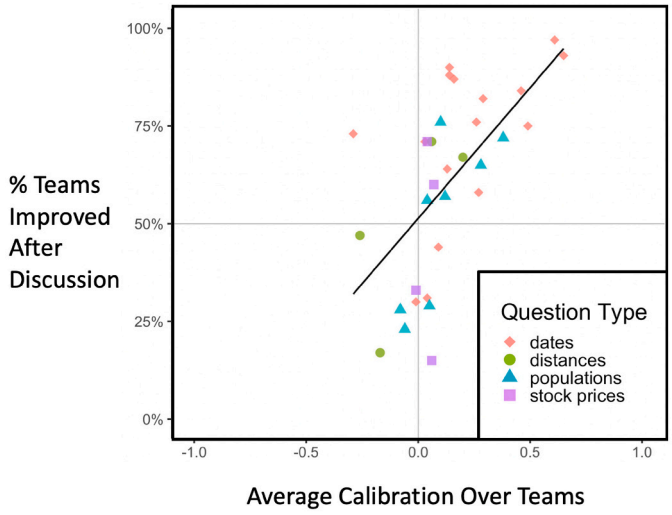

Fig. 3. Data points represent the 32 estimation questions in Studies 1 and 2. For questions with higher average calibration scores prior to discussion, groups were likelier become more accurate after discussion.

improved estimates, but sometimes made them worse.[12] What predicted improvement? In two large studies, we find that a novel, group-level measure of collective confidence calibration predicts whether discussions result in greater accuracy. Groups typically *assume* their most confident members are their most knowledgeable, so whether confidence is *actually* associated with accurate judgment matters. In line with this account, we find that well-calibrated groups were better at recognizing who in their ranks possessed accurate judgment.

As this was the first investigation of collective calibration, we note a few important limitations and potential areas for future investigation. First, our results indicate a robust correlation – not yet a definitive causal link – between pre-discussion calibration and post-discussion improvement. Nonetheless, there are reasons to believe that calibration plays an important role. For one thing, calibration was measured prior to improvement, ruling out reverse causality. For another, the observed relationships were robust across questions, incentives, metrics of improvement, and levels of analysis. It is difficult to identify another variable that might fully explain these patterns. Indeed, after controlling for a host of additional factors, we find a stable link between collective confidence calibration and group improvement (see Appendix).

Still, although participants were randomly assigned to groups, we cannot rule out the possibility that well-calibrated teams and poorly-calibrated ones might have differed in other ways. It is possible that individuals who display better calibration collectively also possess other traits which predispose them to be good discussants. For example, well-calibrated groups might score higher on collective intelligence factors which facilitate conversational receptivity and egalitarian discussion (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010), or maybe they are just more likable and enjoy working together more. Such variables would not necessarily cause pre-discussion calibration, but they may help to further explain how and why pre-discussion calibration and post-discussion improvement are linked.

Second, we have focused on numerical estimation tasks, which have clear correct answers and desirable statistical properties. Would collective calibration also be associated with more effective interactions in other group contexts like creative brainstorming or contentious political debate? While exploring these questions will likely require broader conceptualizations of accuracy and improvement, it will also afford us a clearer perspective on the importance of collective calibration for group performance. Our view is that calibration is likely to be associated with more productive interactions any time (a) groups are likely to listen to their most confident members, and (b) confidence is reliably associated with domain-specific knowledge or expertise which can lead the group towards better answers. This framework also suggests a natural boundary for our account: If groups pay no special heed to confidence, calibration's link to improvement should be weakened.

A further question for future research concerns whether collective calibration is a stable group property across time and situation. Our data contain repeated observations of groups wrestling with consecutive problems of the same type, so we can glean some early insight into calibration's stability by testing whether there is clustering of calibration scores by group. To investigate, we regressed pre-discussion calibration scores on dummy variables for question and compared models that either did or did not include additional dummies for group. Including group effects in the model predicts an additional 20% of the variance in calibration scores, and also increases measures of adjusted-R^2, which penalizes for overfitting. Such results provide suggestive evidence that calibration may be partly stable, at least across consecutive questions of the same type. More work is needed to explore whether calibration generalizes across situations. Still, it would make sense for collective calibration to be at least somewhat trait-like. Well-calibrated teams are comprised of well-calibrated individuals, and individual calibration has been previously linked to stable cognitive traits such as actively open-minded thinking and cognitive reflection (Frederick, 2005; Haran, Ritov, & Mellers, 2013). Building a more complete psychological profile of well-calibrated groups represents an important future direction.

Perhaps the most important follow-up question concerns whether collective calibration can be taught. Can lightweight pre-discussion interventions help groups become better calibrated and if so, would they induce better discussions? Prior researchers have tested, to varying degrees of success, a host of procedures to bootstrap *individual* calibration, including allowing participants to express their rationales for forecasts (Mellers et al., 2014), asking them to generate possible counter-arguments (Hoch, 1985), giving base rates (Mellers & McGraw, 2004), and providing feedback and accountability (Lerner & Tetlock, 1999). Adapting these approaches into efficient training that boosts *collective* calibration may prove challenging, however, because collectively calibrating groups is more complicated than eliminating overconfidence. Any effective intervention to improve collective calibration would need to make some group members (the more accurate) more confident while simultaneously making others (the less accurate) less confident. A key objective for future research should be to test whether teams can be taught to better calibrate their confidence, recognize relative expertise within their ranks, and surface useful knowledge during discussion.

4. Conclusion

People often display exaggerated beliefs about their skills and knowledge. We misunderstand and over-estimate our ability to answer general knowledge questions (Arkes, Christensen, Lai, & Blumer, 1987), save for a rainy day (Berman, Tran, Lynch Jr, & Zauberman, 2016), and resist unhealthy foods (Loewenstein, 1996), to name just a few examples. Such failures of calibration can have serious consequences, hindering our ability to set goals (Kahneman & Lovallo, 1993), make plans (Janis, 1982), and enjoy experiences (Mellers & McGraw, 2004). Here, we show that *collective* calibration also predicts the effectiveness of group discussions. In the context of numeric estimation tasks, poorly calibrated groups were less likely to benefit from working together, and, ultimately, offered less accurate answers. Group interaction is the norm, not the exception. Knowing what we know (and what we don't know) can help predict whether interactions will strengthen or weaken crowd wisdom.

[12] Traces of this "illusion of effective discussion" have been spotted previously (Heath & Gonzalez, 1995; Plous, 1995), and it deserves further attention in future work.

Funding

This research was funded by an NSF grant to the second author (NSF DRMS 15559370), a contract from the Intelligence Advanced Research Projects Activity (IARPA contract 140D0419C0049), and support from the Open Philanthropy Project to the third author. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the U.S. Government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2021.104157.

References

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research, 27*, 123–156.
Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133–144.
Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight: The International Journal of Applied Forecasting, 5*, 3–15.
Asch, S. E. (1955). Opinions and social pressure. *Scientific American, 193*, 31–35.
Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior, 1*, 90–99.
Berman, J. Z., Tran, A. T., Lynch, J. G., Jr., & Zauberman, G. (2016). Expense neglect in forecasting personal finances. *Journal of Marketing Research, 53*, 535–550.
Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559–583.
Dalio, R. (2017). *Principles*. Simon & Schuster.
Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1*, 79.
Einhorn, H., Hogarth, R., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84*, 158–172.
Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25–42.
Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision making, 8*, 188–201.
Heath, C., & Gonzalez, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes, 61*, 305–326.
Hertwig, R. (2012). Tapping into the wisdom of the crowd—With confidence. *Science, 336*, 303–304.
Hoch, S. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 719–773.
Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences, 101*, 16385–16389.
Janis, I. L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascoes* (vol. 349).
Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science, 39*, 17–31.
Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103*, 687.
Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., … Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. In , *113*. *Proceedings of the National Academy of Sciences* (pp. 8777–8782).
Larrick, R. P., Mannes, A. E., Soll, J. B., & Krueger, J. I. (2011). The social psychology of the wisdom of crowds. In *Social Psychology and Decision Making* (pp. 227–242).
Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes, 88*, 605–620.
Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255.
Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*, 272–292.
Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*, 276.
Mellers, B., & McGraw, A. P. (2004). Self-serving beliefs and the pleasure of outcomes. *The Psychology of Economic Decisions, 2*, 31.
Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*, 1106–1115.
Minson, J. A., Mueller, J. S., & Larrick, R. P. (2017). The contingent wisdom of dyads: When discussion enhances vs. undermines the accuracy of collaborative judgments. *Management Science, 64*, 4177–4192.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502.
Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour, 2*, 126.
Plous, S. (1995). Comparison of strategies for reducing interval overconfidence. *Journal of Applied Psychology, 80*, 443–454.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717–731.
Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature, 541*, 532.
Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17*, 39–57.
Raafat, R. M., Chater, N., & Frith, C. (2009). Herding in humans. *Trends in Cognitive Sciences, 13*, 420–428.
Regan-Cirincione, P. (1994). Improving the accuracy of group judgment: A process intervention combining group facilitation, social judgment analysis, and information technology. *Organizational Behavior and Human Decision Processes, 58*, 246–270.
Sah, S., Moore, D. A., & MacCoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes, 121*, 246–255.
Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2012). Why groups perform better than individuals at quantitative judgment tasks: Group-to-individual transfer as an alternative to differential weighting. *Organizational Behavior and Human Decision Processes, 118*, 24–36.
Sherif, M. (1937). An experimental approach to the study of attitudes. *Sociometry, 1*, 90–98.
Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2010). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research, 38*, 1–15.
Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*, 122–124.
Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*, 1467.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes businesses, economies, societies, and nations*.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686–688.
Zarnoth, P., & Sniezek, J. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33*, 345–366.