How expertise mediates the effects of numerical and textual communication on individual and collective accuracy

Nicholas Beauchamp^{1,*}, **Sarah Shugars**^{2}, **Briony Swire-Thompson**^{1,3}, **and David Lazer**^{1,3}

^{1}Network Science Institute, Northeastern University, Boston, MA 02115; ^{2}School of Communication and Information, Rutgers University, New Brunswick, NJ 08901; ^{3}Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138; *To whom correspondence should be addressed. E-mail: n.beauchamp@northeastern.edu

Performance on difficult tasks such as forecasting generally benefits from the "wisdom of crowds," but communication among individuals can harm performance by reducing independent information. Collective accuracy can be improved by weighting by expertise, but it may also be naturally improved within communicating groups by the tendency of experts to be more resistant to peer information, effectively upweighting their contributions. To elucidate precisely how experts resist peer information, and the downstream effects of that on individual and collective accuracy, we construct a set of event-prediction challenges and randomize the exchange of both numerical and textual information among individuals. This allows us to estimate a continuous nonlinear response function connecting signals and predictions, which we show is consistent with a novel Bayesian updating framework which unifies the tendencies of experts to discount all peer information, as well as information more distant from their priors. We show via our textual treatment that experts are similarly less responsive to textual information, where non-experts are more affected and benefited overall, but experts are helped by the highest-quality text. We apply our Bayesian framework to show that the collective benefits of expert nonresponsivity are highly sensitive to the variance in expertise, but that individual predictions can be "corrected" back towards their unobserved pre-treatment states, boosting the collective accuracy of non-experts close to the level of experts, and restoring much of the accuracy lost due to intra-group communication. We conclude by examining potential avenues for further improving collective accuracy by structuring communication within groups.

Collective intelligence | Forecasting | Social Influence | Expertise | Reasoning | Cognitive biases

ACKNOWLEDGMENTS. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17061500006 and 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Introduction

The superiority of expertise was somewhat diminished a century ago by the discovery that an aggregate of non-experts can, in many circumstances, outperform even the most skilled individual via the “wisdom of the crowd” (Galton, 1907; Hong and Page, 2004; Page, 2018). But the importance of expertise has undergone a recent resurgence: On the one hand, giving more weight to the contributions of experts can outperform a generic crowd (Mannes et al., 2014; Budescu and Chen, 2015), and in some cases only giving weight to the top experts can be optimal (Mellers et al., 2015; Tetlock and Gardner, 2016). On the other hand, a crowd often has weaknesses in real-world settings: in particular, individuals tend to communicate extensively within groups (Yaniv, 2004), and this social influence may reduce collective intelligence in various ways. Shared information can induce copying and thereby reduce the amount of independent information in a group (Surowiecki, 2004; Golub and Jackson, 2010; Lorenz et al., 2011; Sunstein and Hastie, 2015; Page, 2018; Sommers, 2006; Patry, 2008; Mercier and Landemore, 2012; Yaniv, 2004; Sunstein and Hastie, 2015); it can lead to opinion convergence toward suboptimal solutions due to information cascades, particularly in smaller groups (King and Cowlishaw, 2007a; Golub and Jackson, 2010; Lorenz et al., 2011); and it can reduce overall diversity of opinion (Hong and Page, 2004; Page, 2018).

But alongside the rise in the importance of expertise has also come increasing awareness that identifying experts for novel or complex real-world tasks can be a major challenge (Madirolas and de Polavieja, 2015; Tetlock and Gardner, 2016; Luo et al., 2018; Mannes et al., 2014; Budescu and Chen, 2015), while self-assessment is notoriously unreliable (Kruger and Dunning, 1999; Dunning et al., 2003; Schlösser et al., 2013; Atir et al., 2015; Madirolas and de Polavieja, 2015). Nor are the drawbacks of communication necessarily as large as once thought: providing participants with better information can boost collective accuracy (King and Cowlishaw, 2007b; King et al., 2011; Jayles et al., 2017; Luo et al., 2018; Toyokawa et al., 2019), in part by reducing sensitivity to misleading information or outlier opinions (King et al., 2012; Kao et al., 2018). Furthermore, the harm from diversity reduction may not be as great as had been supposed (Jayles et al., 2017; Nobre and Fontanari, 2020). Indeed, recent work has suggested that under some circumstances, one can have the best of both expertise and communicating crowds: since experts tend to be more resistant to social influence (Tversky and Kahneman, 1974; Yaniv, 2004; Kaustia et al., 2008; Welsh et al., 2014; Madirolas and de Polavieja, 2015; Cheek and Norem, 2017; Luo et al., 2018; Jayles et al., 2017), this resistance can be used as a potential proxy for expertise, and may even serve as a natural upweighting of expert opinion within communicating crowds due to non-experts being more persuaded by experts than vice versa (Becker et al., 2017).

However, the circumstances under which expertise can naturally boost the collective accuracy of a communicating crowd remain ill-defined, in part because the mechanisms by which experts resist persuasion remain poorly understood. In order to better understand how communication and expertise interact to affect the wisdom of the crowd, we constructed an experiment using a set of complex real-world event prediction problems, and manipulated the exchange of both numerical and textual information between subjects. Our experiment was a 2x2 within-subject design with two independently randomized treatments: with 50% probability per prediction task, a subject was shown a single randomly-selected numeric prediction made by one of their peers prior to making their own prediction; and independently, with 50% probability per task, a subject was shown a peer's written reason justifying their prediction. The effects of these treatments, and the interaction between those effects and expertise, were then assessed on prediction values, on prediction accuracy, and on aggregate accuracy.

Our design allows us to replicate and synthesize a number of previous findings, as well as demonstrating a number of new ones. Our continuous-valued prediction task allows us to measure the precise response curve relating peer information to the predictions made by receivers. This response curve shows the previously-established diminished responsivity of experts, as well as their tendency to more aggressively discount information in proportion to its distance from their prior belief. We present a new model that combines both of these effects in a single Bayesian framework, and show that the theoretical response curves closely match the empirically derived ones. We use this framework to show via an analytic example and also a simulation how group accuracy can be either helped or harmed by communication. We show that the benefits and harms are highly sensitive to the variance in expert responsivity, to the size of the group, and to the number of exchanges in the group. We then use our Bayesian framework to develop a new method to "correct" predictions affected by peer information, shifting them back towards the (entirely unobserved) prior predictions. We find that this correction benefits non-expert groups more than expert groups, that it is comparable to benefits from traditional reweighting procedures, and that it raises the accuracy of non-expert groups to the level of relative experts, primarily by undoing the damages to group accuracy due to intra-group communication.

Our text treatment shows for the first time that the effects of the most common form of communication, linguistic, are similar to those for numeric information, including the sensitivity of those effects to the expertise of the receiver. It is often assumed that experts discount more distant information due to its lower quality. However, with numeric information it is impossible to distinguish a generic bias towards one's prior from a judgment based on the inherent quality of a piece of information, since those two are correlated. Our textual treatments allow us to measure information quality directly, both via peer judgments, and via features such as containing numbers or URLs. We find that non-experts are helped by peer text more than experts, but that experts are indeed helped, but only by the highest-quality information. Thus experts appear to be both more confident in their own beliefs, and more likely to discount low-quality numerical or textual information. These two biases can benefit not just experts individually, but also the group as a whole, albeit in ways that are highly sensitive to group size, quantity of communication, and variance in expertise.

Taken as a whole, we provide a new, unified Bayesian framework incorporating these two expertise effects; show for the first time that these effects operate for text as well as numbers, and that the distance-based effect is likely due to information quality; and demonstrate how these effects may or may not benefit the group, as well as how they can been boosted by "correcting" peer-affected predictions after the fact. Our results suggest new avenues for shaping the flow of peer information to boost the wisdom of the crowd, as we discuss in the conclusion.

Methods

To evaluate the impact of both numerical and textual social influence on individuals and on collective intelligence, we constructed a series of real-world event-prediction tasks. Each subject was presented with up to 16 questions about future events randomly ordered, with four questions in each of four topic areas: politics, entertainment, economics, and natural events such as diseases or weather. Questions were brief and on continuous scales, such as "What will be the approval rate for the Russian government at the end of January" or "What will be the value of one bitcoin in USD at 11:59pm on the 21st of January" (see Table S6 for the full list). Each prediction event resolved between five days and five weeks into the future. Subjects were asked to provide their best numerical prediction for each question; provide subjective judgments of their confidence in their prediction and expertise on the topic; and 50% of the time they were also asked to provide a brief textual justification or "reason" for their prediction. Due to the voluntary nature of most of our subjects, subjects were not required to answer all questions, with an average of 7.8 questions answered per subject. After making their predictions, they were asked a series of demographic questions, a six question political knowledge quiz, and a three question Cognitive Reflection Test (Frederick, 2005) to measure basic reasoning ability.

Our experiment was a 2x2 within-subject design with two independently randomized treatments: with 50% probability per prediction task, a subject was shown a single randomly-selected numeric prediction made by one of their peers prior to making their own prediction (treatment) or not (control); and independently, with 50% probability per task, a subject was shown a peer's written reason justifying their prediction (treatment) or not (control). Reasons were delivered independent of the numeric predictions, but were almost always written in a way that could be sensibly interpreted even without an associated number. Numeric treatments were selected by drawing random values from a uniform distribution covering all but the highest and lowest 5% of existing predictions; this was designed to omit non-compliant predictions while preserving the mean and variance of the distribution, and to also provide a higher density of extreme predictions in order to better measure tapering effects as discussed below. Unlike some previous work, we did not elicit predictions prior to presenting the treatment, which is consistent with how almost all real-world peer interactions occur; as we will show, pre-elicitation is not necessary for assessing treatment effects. No identifying characteristics of the sender were included with the treatment, and there was no mechanism for interaction or repeated exchanges, merely the one-time delivery of a number and/or piece of short text.

Our primary outcomes were the individual's prediction value; the accuracy of that prediction as measured by the squared error between a prediction and the true outcome after standardizing by question; and the squared error of the per-question mean prediction of various subsets of subjects, such as those with higher or lower expertise. For textual treatments, we constructed a number of measures of "quality," using both natural language processing (NLP) methods, and by asking a randomly selected subset of recipients to rate the quality of the text they read and aggregating those ratings for each text item.

Our study had 804 unique participants: 494 unpaid volunteers recruited from Reddit and Craigslist, and 310 paid Amazon Mechanical Turk workers. All participants provided informed consent, and our study was approved by the IRB at Northeastern University (IRB # 13-03-09). 84% of predictions were made by volunteers, and there was no statistically significant difference in accuracy between the two subject pools, nor any significant interactions between pool type and treatment effects (Table S5). The final sample was 65% male with a mean age of 33 ($sd = 10$). Prior to analysis, the top and bottom 1% outliers were removed as being mainly non-compliant, and predictions were standardized by question to mean 0 and standard deviation 1 before all analysis; prediction values were not logged because distributions were approximately normal in most cases (Lorenz et al., 2011; Luo et al., 2018; Kao et al., 2018).

Our key mediating variable is subject expertise. "Expertise" can encompass many possible concepts and has been operationalized into several measures in the literature (Tversky and Kahneman, 1974; Dunning et al., 2003; Becker et al., 2017; Guilbeault and Centola, 2020; Attali et al., 2020). Following the literature, our core definition of "expertise" is domain-specific skill. To capture this direct measure of ability, we assessed each subject's accuracy in the first 50% of the (randomly ordered) prediction questions, which allows us to use this measure of expertise as a covariate in the second half of responses.

We also tested a number of alternative measures of expertise. To measure self-judged ability, we asked subjects their degree of confidence in each question as well as their self-assessed "expertise" in that topic area. And as additional objective measures, we included a three-question reasoning quiz, education level question, and a political knowledge test. As other have found, we observed that self-judged "expertise" was in fact negatively correlated with accuracy ($\beta = -0.17, p < 0.001$; Table S2:M1), possibly due to the Dunning-Kruger effect (Dunning et al., 2003; Schlösser et al., 2013; Nuhfer et al., 2016, 2017). And while "confidence" was significantly associated with accuracy ($\beta = 0.11, p < 0.001$, Table S2:M1), this effect disappears when including question-level random effects, suggesting that "confidence" measures a question-level quality (such as the ease of the problem) rather than a subject-level quality (such as expertise or skill). Similarly, neither education nor political knowledge were associated with prediction accuracy when controlling for other features (Table S5). However, both reasoning ability and our domain-specific measure of "expertise" were significantly predictive of accuracy in the second half of questions answered (Table S2, Models 4 vs 7). We hereafter use this domain-specific accuracy measure as our core expertise measure, but also discuss reasoning ability in places.

Results

1. Peer Effects on Individual Predictions.

Numeric Treatments. Previous work has theorized that responsivity to peer information varies with both receiver expertise, and the distance between the receiver's original prediction and the peer prediction that they are shown

(Yaniv, 2004; Kaustia et al., 2008; Welsh et al., 2014; Madirolas and de Polavieja, 2015; Cheek and Norem, 2017; Luo et al., 2018; Jayles et al., 2017; Becker et al., 2017). Most previous theoretical models of this process have followed some variant of the DeGroot framework, where receivers react to information in a mechanistic way ungrounded in any specific Bayesian information processing framework. While the DeGroot model has the advantage of generality, being consistent with many different forms of information updating, these models lack specificity regarding precisely how the receiver's responsivity varies with expertise and the observed signal. We instead propose here a simple Bayesian updating framework, which incorporates in a single functional form the reduced responsivity of experts to all signals; the reduced responsivity of every receiver to signals more distant from the receiver's belief; and the tendency for experts to discount distant signals more severely than non-experts. Our model does not provide a rational basis for either the reduced responsivity of experts or the reduced responsivity with distance; rather, it illustrates how these two empirically-demonstrated psychological tendencies may operate within a broader Bayesian updating framework.

In the standard Bayesian framework, if a receiver has a normal prior belief about the truth μ with expected value α and uncertainty (variance) σ_1^2, and observes a signal β to which they attribute normally distributed uncertainty σ_2, their posterior belief will be:

$$f(\mu|\alpha,\beta) \propto \frac{1}{\sqrt{\sigma_1^2\sigma_2^2}} exp\{-\frac{(\mu-\alpha)^2}{2\sigma_1^2} - \frac{(\mu-\beta)^2}{2\sigma_2^2}\} \tag{1}$$

After some algebraic manipulation, the receiver's updated expectation regarding the position of the truth μ is therefore the average of α and β, weighted by σ_1^2 and σ_2^2:

$$E[\mu|\alpha,\beta] = \frac{\sigma_2^2\alpha + \sigma_1^2\beta}{\sigma_1^2 + \sigma_2^2} \tag{2}$$

The uncertainty assigned by the receiver to their prior belief α, σ_1^2, and the uncertainty assigned by them to the observed signal β, σ_2^2, can account for different aspects of expert responsivity. (1) To account for the reduced overall responsivity of experts to signals, we hypothesize that experts are more resistant to external signals because they are more confident in their own prior: ie, they will have a relatively lower σ_1^2. (2) To account for the reduced responsivity of everyone to more distant signals (which is empirically motivated and may not be rational), we hypothesize that σ_2^2, the uncertainty or discounting assigned to the observed signal β, is proportional to the distance between α and β; most simply, $\sigma_2^2 = \gamma(\alpha-\beta)^2$. This yields:

$$E[\mu|\alpha,\beta] = \frac{\gamma(\alpha-\beta)^2\alpha + \sigma_1^2\beta}{\sigma_1^2 + \gamma(\alpha-\beta)^2} \tag{3}$$

(3) Since γ is just a proportionality constant, we can also account for experts discounting distant signals more aggressively if γ is relatively higher for experts. Two illustrative examples of this response function are shown in Figure 1 (inset): the expert's response curve (purple) has lower σ_1^2 and higher γ, while the non-expert curve (blue) has higher σ_1^2 and lower γ.

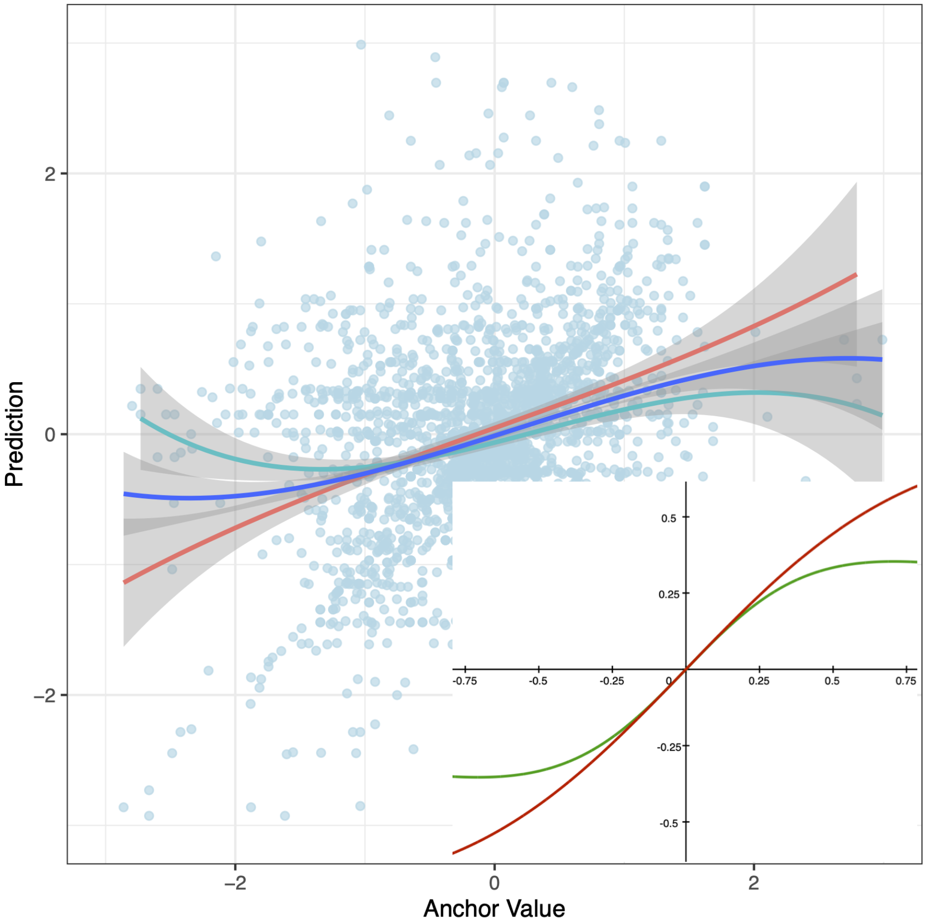

Fig. 1. Main: Response curves of subject predictions as a function of peer information signals. Blue dots: all subjects. Steep red line: cubic fit for low-expertise subjects; middle blue line: all subjects; flatter green line: high-expertise subjects. Inset: Illustrative response curves based on equation 3, where $\alpha = 0$ and $\{\sigma_1^2, \gamma\}$ equal to {1,2} (green, experts) vs {2,1} (red, non-experts).

To test how well our empirical data match the response function postulated in Equation 3, we begin with a pooled model: after standardizing all predictions by question, we regress each subject's post-treatment prediction on their received peer prediction value, using a cubic polynomial to capture the tapering effect of signals with distance. Although some previous work has assumed it to be necessary to measure pre-treatment positions in order to estimate treatment effects (Becker et al., 2017, 2019; Jayles et al., 2017; Nobre and Fontanari, 2020; King et al., 2012; Kao et al., 2018; King and Cowlishaw, 2007b; King et al., 2011; Jayles et al., 2017; Luo et al., 2018; Toyokawa et al., 2019), this is in fact not necessary: as Figure 1 (main, blue line) shows clearly, predictions vary linearly with signals with tapering effects for the most distant signals. In the cubic polynomial regression, we find the linear treatment effect is approximately one third the anchor value ($\beta = 0.33, p < 0.001$; Table S1:M1), i.e. for each standard deviation increase in anchor size, the prediction is increased by one third of a standard deviation. This is consistent with previous work (Mussweiler and Strack, 2000; Yaniv, 2004; Mavrodiev et al., 2013; Cheek and Norem, 2017; Jayles et al., 2017). Also consistent with that work, we find that this linear effect tapers off with distance (cubic $\beta = -0.02$, $p = 0.015$; Table

S1:M1), with the maximum effect topping out at approximately 2/3 of a standard deviation from the group mean (Kao et al., 2018).

As Figure 1 (inset) shows, the functional form in Figure 1 (main) closely resembles our Bayesian functional form. To test whether our empirical response function varies with expertise, we divide our pool into higher-expertise "experts" and lower-expertise "non-experts" and estimate models separately for these two groups. To capture our domain-specific measure of expertise, we assess subject performance in the first 50% of questions answered, and then estimate the polynomial model separately for the high and low groups using the second 50% of questions answered. Figure 1 (main) shows the curves for non-experts (red) and experts (green), closely resembling the functional forms illustrated in the inset. The polynomial regression shows that the linear persuasive effect is considerably stronger for the low-expertise group than the high (high $\beta = 0.230$; low $\beta = 0.352$; both $p < 0.001$), and that the tapering effect is significant only for the high-expertise forecasters (cubic effect for high, $p = 0.027$; for low $p = 0.360$).

Because this dichotomization analysis reduces the available information inherent in our continuous measure of expertise and makes statistical tests of difference difficult (MacCallum et al., 2002), we additionally interact expertise with anchor values, using only data from the second 50% of questions answered. We find a significant negative interaction ($\beta = -0.113$, $p < 0.001$; Table S1:M5). This provides additional statistical support for the theory that the effect of the anchor value diminishes with expertise. These results are consistent with previous findings that experts are less affected by information overall and that everyone is less affected by information that is inconsistent with their existing beliefs. However, our continuous-valued treatments allow us to measure precise empirical response curves for different levels of expertise. These empirical response curves closely match the curves generated by our Bayesian model in Equation 3, where experts have lower σ_1^2 (greater self-confidence) and higher γ (higher discounting with distance).

Text Treatments. Although the vast proportion of human communication is linguistic or textual, most previous work on the mediating role of expertise has focused on purely numerical communication (Yaniv, 2004; Kaustia et al., 2008; Welsh et al., 2014; Madirolas and de Polavieja, 2015; Cheek and Norem, 2017; Luo et al., 2018; Jayles et al., 2017), or at most, stylized categorical signals (Guilbeault et al., 2018; Becker et al., 2019). We have shown that experts are less affected by numerical peer information because they are more confident in their own beliefs overall, and also that they more steeply discount inconsistent signals. Although we have fit these two psychological features into a broadly Bayesian framework, this doesn't explain the origins of these tendencies. It is often hypothesized that experts are less affected by peer information because they are more sensitive to information quality, but with numerical information, it is impossible to distinguish quality from distance, since experts' prior beliefs will in general be closer to the truth. However, text allows us to measure quality directly, and directly determine the role that information quality plays in expert discounting of peer information.

Our first numerical result is that experts are in general less affected by peer information. To test this with our text treatments, we hypothesized that the effect of a reason should be similar to the effect of the numerical prediction made by the same individual who wrote the reason, even if the numerical prediction is not seen: a high prediction will have a reason arguing for a relatively high value, and vice versa. If textual effects are not just coarsely in the same direction as the anchor but also of similar magnitude, then by analogy with numeric effects, we would expect the persuasive effect of the latent, unseen anchor value to be weaker for more expert subjects. To test this, we first regressed each receiver's prediction on the unseen prediction made by the writer of the text they viewed, and found a significant positive effect ($\beta = 0.07$, $p = 0.002$; Table S4), about one quarter the size of seeing a numerical prediction directly. This effect was also not driven by the prediction value itself being included in the reason: only 6% of reasons included the literal prediction, and results are unchanged when those are excluded. This result in itself is notable, as there have been few studies to our knowledge to show quantitative variation in textual effects that are well-predicted by unrevealed features of the writer (here, the writer's numerical prediction).

To test whether this text effect is moderated by expertise, we again subdivided by high and low expertise subjects, and found that the effect holds only for low-expertise subjects (high: $\beta = 0.06$, $p = 0.197$; low: $\beta = 0.15$, $p = 0.002$). Similarly, interacting expertise with the (unseen) text-associated anchor value also shows a significant negative interaction ($\beta = -0.064$, $p = 0.04$). Thus low-expertise users are more affected by reasons, just as they are more affected by numerical signals.

In addition to the overall lower effect on experts, our numerical model also showed that experts discount more distant signals more aggressively than non-experts. One hypothesis is that this is because experts are better able to judge information quality. Although with numerical signals we cannot distinguish information quality from distance, with text we can measure information quality directly. To do this, we asked a subset of receivers to rate the quality of the reasons they read. Since using the receiver's own rating as a predictor would open the door to misleading reverse causation (eg, if more accurate subjects tend to rate reasons more highly), we instead use for each reason the mean rating of all subjects who saw and rated that reason. Since on average only 2-3 subjects rated each reason, we include all ratings in the analysis below, but if we exclude the target receiver's rating, our results are slightly weaker but substantively the same. The quality of reasons varied widely, though even the worst-rated reasons were generally compliant and well-intended. Highly-rated reasons often contained URLs or numbers, such as "https://earthquake.usgs.gov/earthquakes/browse/stats.php provides statistic about earthquakes" or "The Shape of Water was nominated 12 times, compared to previous years of La La Land which was nominated 11 times and won 5 times." By contrast, the lowest-rated reasons were generally less informative, eg "Just a guess since I have no idea regarding the politics of the country" or "Trump real clear politics may have both positive and negative feedbacks." Regressing predictions made by those treated with a reason on the interaction between the unseen reason-associated anchor value and the quality score, we find that only higher-quality reasons have a persuasive effect. Dichotomizing by expertise, though, we find that this only holds for high-expertise subjects (Table S4, Models 6-7). For experts, low quality reasons have little to no effect while high-quality reasons have effects in the expected direction, whereas for non-experts there is no differential effect. A three-way interaction (Table S4:M8) also shows the same moderation, where higher ratings and expertise intensify the effects of a reason. As a robustness check, we also reran these analyses removing the target receiver's own rating from the mean, and our results, though slightly weakened due to removing many ratings, were substantively the same. So just as with numerical communication, low-expertise subjects are broadly affected by all forms of textual communication, but high-expertise subjects are much more selective, and only affected by high-quality information. Our results show that this selectivity seems to be a general feature of expertise, holding for both numerical and textual communication, and our textual results provide evidence that this selectivity is indeed due to greater sensitivity to information quality.

2. Peer Effects on Individual Accuracy.

While we have established how expertise mediates the effects of peer communication on the receiver's predictions, of more practical importance is the effect on the accuracy of those predictions. However, the effects of communication on prediction values do not directly translate to accuracy: because lower-expertise subjects are in general farther from both the truth and the group mean than high-expertise subjects, a random exchange of predictions should benefit individual low-expertise subjects more than higher-expertise subjects. However, if experts are more resistant to low-quality information, the higher quality information they do consider may result in comparably higher accuracy for those subjects.

To determine how expertise mediates the effects of numerical and textual information on accuracy, we first regress accuracy (squared error, reverse-coded hereafter so that higher is better) on anchor and reason treatments, question order, confidence, self-assessed expertise, question order and time taken (Table S2). We find that in general individuals do benefit from seeing a peer's a prediction ($p < 0.001$), and when we split our sample by higher- and lower-expertise users (Figure 2), the low-expertise (red) users benefit slightly more from seeing a peer's anchor value than the high-expertise (blue). Similarly, when it comes to seeing a peer's reason, the lower-expertise subjects again benefit more than the more expert subjects (Table S2: low: $\beta = 0.40$, $p = 0.007$; high: $\beta = 0.10$, $p = 0.475$). When we interact either expertise or reasoning ability with the numerical and textual treatments, all four interactions are negative, although each is below traditional statistical significance levels (see Table S2). Considered as a whole, these results support previous work showing that the benefits of peer information diminish with expertise (Yaniv, 2004; Kaustia et al., 2008; Welsh et al., 2014; Madirolas and de Polavieja, 2015; Cheek and Norem, 2017; Luo et al., 2018; Jayles et al., 2017).

This overall decline in benefit with expertise, however, does not answer whether this is due to experts being particularly resistant to poor information, or simply due to the fact that those with less accurate predictions will in expectation benefit more from communication even when everyone is equally receptive. But while we have no way of measuring numerical information quality independent of accuracy, we can again directly measure textual information quality. Using our peer-rating of reason quality, we found that the higher the mean reason rating was

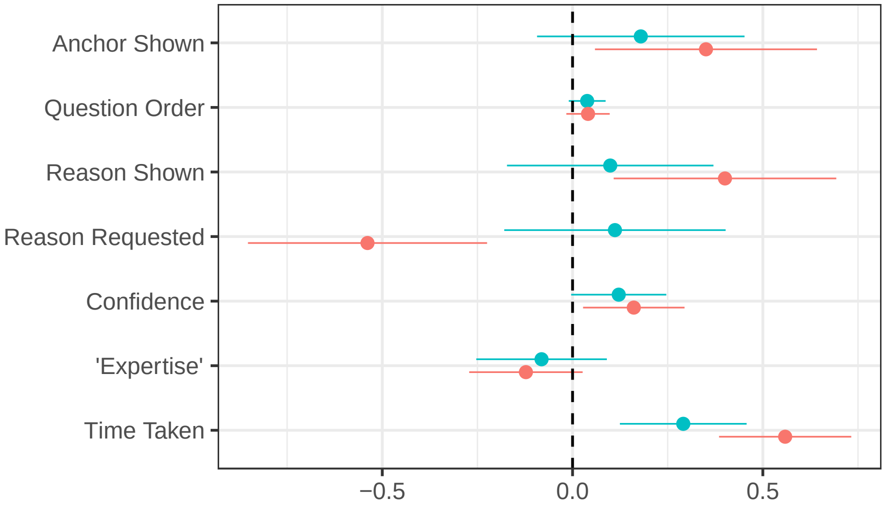

Fig. 2. OLS regression of individual accuracy on treatments and other contributory factors using high/low expertise split sample, where expertise is determined from the first 50% of predictions made and the model is fit on the second 50%. Top, blue: High-expertise. Bottom, red: Low-expertise. ‘Expertise’ is self-judged; ‘Reason Shown’ is whether a subject saw a reason, while ‘Reason Requested’ is whether they were asked to provide a reason (see Table S2).

for those who received a reason, the higher the accuracy of the receiver’s prediction ($\beta = 0.08$, $p = 0.05$; Table 1). This is an important finding in its own right: higher-quality textual information boosts the accuracy of those who receive it. But when we divide our sample by expertise, we find that this relation only holds for the higher-expertise subjects (high: $\beta = 0.21$, $p = 0.046$; low: $\beta = 0.01$, $p = 0.382$). Interacting expertise with reason quality shows similar results: the interaction is significantly positive (interaction $\beta = 0.19$, $p = 0.02$), indicating that the benefits of reason quality are strongest for the most expert receivers. As a robustness check, excluding the target receiver’s own rating from the aggregate reason rating somewhat weakens these results: the signs remain the same, but the effects fall below traditional significance thresholds. More data would be needed to clarify whether this is due to reduced N, or some degree of reverse causation where high-expertise subjects are indeed somewhat more likely to rate reasons more highly. Overall, these results suggest that experts are more selective in distinguishing good from bad information, rather than everyone being affected equally and the less accurate users simply benefiting more in expectation.

In order to glean some insight into what specific features of these free-text treatments are responsible for their benefits, we constructed a number of automated measures of reason quality: word count; average word length; count of number use; count of all-capitalized words; count of questions; and count of URLs. Of these, URLs and numbers appear to confer systematic benefits, presumably because they contain concrete and usable information (Table S3:M1). When splitting by expertise, only numbers are significant, and only for experts (Table S3, Models 2-3). When interacting the significant features with expertise, the only significant interaction is once again with numbers, where the interaction is positive: the more expert the user, the more they are benefited by numbers shared within a reason (Table S3:M4). Note that only 6% of the reasons which had a number included the actual unseen prediction.

Table 1. Effects of reason's user-rated quality on accuracy

	Dependent variable:			
	MSE Accuracy			
	All	High	Low	Int
Reason Quality	0.084^{*}	0.271^{**}	0.007	0.337^{***}
	(0.051)	(0.112)	(0.093)	(0.110)
Expertise				0.094
				(0.083)
Expertise * R. Quality				0.189^{**}
				(0.080)
Constant	-1.093^{***}	-1.081^{***}	-0.946^{***}	-0.917^{***}
	(0.051)	(0.115)	(0.098)	(0.114)
Observations	2,204	404	440	844
R^2	0.001	0.014	0.00001	0.013
Adjusted R^2	0.001	0.012	-0.002	0.009
Note:				*p<0.1; **p<0.05; ***p<0.01

These results are robust to excluding those observations. Of the URLs, some pointed to websites with concretely useful information such as prices, while others pointed to more general websites such as CNN. None pointed to direct answers since our prediction tasks were chosen to not have knowable solutions. This may in part explain why the indirect information in web pages—which would need to be interpreted—may benefit experts more than non-experts. There are many further ways to extend natural language processing (NLP) techniques to determine which features are associated with higher-rated reasons, but our findings here indicate that in general experts are benefited only by higher-quality information, either as qualitatively rated by peers, or containing concretely useful information like numbers or websites.

3. Peer Effects on Collective Accuracy.

If effects on individual accuracy are arguably of more practical importance than direct effects on predictions, of even greater practical importance are the effects on group accuracy. Numerous studies have shown that aggregating individual predictions can out-perform the accuracy of even the most expert individuals, so the mediating effect of expertise is most practically important for aggregate predictions.

The fickle benefits of expertise. Although we found that seeing a peer prediction or a peer's reason increases the receiver's individual accuracy on average, we found no significant benefits for group accuracy. We had 16 questions each with equal numbers of treated and untreated respondents, so to test for differences in group accuracy we averaged the predictions in each of the 2x16=32 groups, calculated the squared error between a group's prediction and the true outcome, and then conducted a paired T test between the 16 treated group errors and 16 untreated group errors. The p value from that test for the numeric treatment was 0.84, while the p value for the text treatment was 0.26.

Since these tests involve only 16 paired observations, the statistical power to detect small differences is weak, and thus this non-result is not in itself surprising. However, there are also well-known substantive reasons why information exchanges that help the average individual may nevertheless harm, or leave unchanged, collective accuracy. Peer communication has sometimes been considered harmful to collective accuracy because it reduces group diversity (Page, 2018) by reducing individual variance. The "collective error" or error of the group mean, $(\bar{x} - \mu)^2$, is mathematically equal to the "average individual error" $\frac{1}{n}\sum(x_i - \mu)^2$ minus the "group diversity" $\frac{1}{n}\sum(x_i - \bar{x})^2$. It was initially thought that reducing group diversity via peer communication would therefore increase the collective error (Lorenz et al., 2011), but this is now known to be mistaken both empirically and theoretically (Becker et al., 2017; Nobre and Fontanari, 2020): reducing diversity via interpersonal communication will generally also lower the average individual error, and thus would mathematically have no net effect on collective accuracy.

But this line of reasoning is altered when the effects of communication are asymmetrical with respect to individual expertise. It is well know that aggregate predictions such as the weighted mean, when weighted by expertise, can outperform simple averages (Mannes et al., 2014; Budescu and Chen, 2015; Mellers et al., 2015; Tetlock and Gardner, 2016; Jayles et al., 2017). But signals whose effects vary with the expertise of the receiver are equivalent to weighting by expertise, as was shown in Equation 2: experts will weight their own predictions more, and thus the average of experts and non-experts who have exchanged peer information will weight the expert predictions more. Thus if low-expertise subjects are affected by communication more than high-expertise, this has a similar effect on collective accuracy as directly weighting experts when constructing an aggregate – but without need to directly weight, or indeed to even measure, expertise.

Previous work has empirically shown that decentralized groups can benefit from this implicit reweighting (Becker et al., 2017). However, these benefits appear highly sensitive to group size, initial conditions, and information cascades (Toyokawa et al., 2019). To illustrate how the benefit to the group is sensitive to expertise responsivity even in the absence of cascades, consider a simple example with two players. One is an expert with initial prediction equal to 0 (which is the truth, unbeknownst to either player), and the other is a non-expert with initial prediction equal to 2. The "average individual error," ie the mean of their two individual squared errors, is $(0^2 + 2^2)/2 = 2$, while the "collective error," or the squared error of the group mean, is 1; thus as usual, the group mean is more accurate than their predictions considered individually. If they both see the other's prediction and update their own prediction to $\gamma *$own + $(1 - \gamma)$*other, the mean of the two predictions remains the same regardless of γ, and thus the error of the group mean remains the same. However, consider two cases where there is a single exchange: in the first case the expert and non-expert are equally responsive, while in the second case the expert is less responsive. In case (1), say that the mutual influence is such that the expert updates to 1 if they see the non-expert's prediction, and likewise for the non-expert. Then the squared error of the group mean if the non-expert sees the expert is 1/4, the squared error if the expert sees the non-expert is 9/4, and the expected value is therefore 10/8, or worse than no information exchanged (=1). But in case (2), say that the non-expert updates as before, but the expert only updates half as much, to 0.5 upon seeing the non-expert's prediction of 2. In that case, the squared error if the non-expert sees the expert is 1/4, the squared error if the expert sees the non-expert is 25/16, and the expected value is 29/32 – i.e., better than no information exchanged (=1). Thus whether the group is benefited or harmed depends on group size, number of exchanges, the distribution of initial beliefs, and most importantly, the relative responsivity of experts vs non-experts.

To examine these tradeoffs in a more realistic setting, we constructed a simple simulation within the context of our Bayesian framework. In this simulation, we varied group size (N), the degree to which persuasive effects vary with expertise (Exp), and the number of exchanges among between members. Each member receives a permanent "expertise" level and initial prior belief, where the higher the expertise, the closer the initial belief is to the truth, and also the lower σ_1^2 and higher γ are for that individual. $Exp = 0$ means that all members have the same σ_1^2 and γ, while higher Exp values mean a wider range of variation of individual expertise levels, with some agents having low σ^2 and high γ (more expert) while others the reverse (less expert). Members then sequentially and randomly receive signals from each other, updating each time according to Equation 3 (see SI for details).

Figure 3 shows the difference in collective accuracy between the pre- and post-communication stages for various group sizes N, variance in expert responsivity Exp, and numbers of exchanges. Values are differences in group MSE between pre and post, normalized by the mean MSE to give a sense of the percentage improvement; values above 0 show an improvement in accuracy between pre and post, while those below 0 show a decline. When there is no expertise-dependent sensitivity and everyone responds equally ($Exp = 0$), peer communication harms collective accuracy relative to the no-communication baseline (red line). But when we increase the variance in responsivity of experts to peer information (Exp = 0.5), communication improves accuracy over the baseline. Furthermore, increasing the number of exchanges intensifies these effects: shifting from N total exchanges per group (top) to 10N exchanges (bottom) increases both the harm and the help from information exchange.

In our experimental setting, the number of individuals per group (prediction problem) is relatively high, the difference in responsivity between experts and non-experts is relatively small, and the number of exchanges is on the order of N, all of which put us closer to the no-help regime. Additionally, the number of our unique groups is small at 16: running our simulation 16 times (instead of 2000, as shown) yields confidence intervals that overlap 0 at all N and Exp values (not shown). Practically, this suggests that in many empirical settings, the gains to collective wisdom due to the reduced responsivity of experts may be only detectable when the group is small, the number of exchanges within each group is high, and there are many groups to analyze. In terms of designing an improved mechanism for boosting group accuracy, in general one would not prefer a smaller group since larger groups are more accurate, but given a fixed group size determined by one's budget, these results suggest that additional gains from expert responsitivty might be found in (1) increasing the number of exchanges per group and (2) increasing the differential expertise effect, by making experts less reponsive and/or non-experts more responsive. (2) could be implemented either indirectly, by highlighting the reputation of successful subjects, or directly, by censoring the information sent from non-experts to experts.

Improving group accuracy. Although the natural boost to collective accuracy produced by the reduced responsivity of experts to peer information may be fragile in many settings, we can nevertheless improve upon nature after the

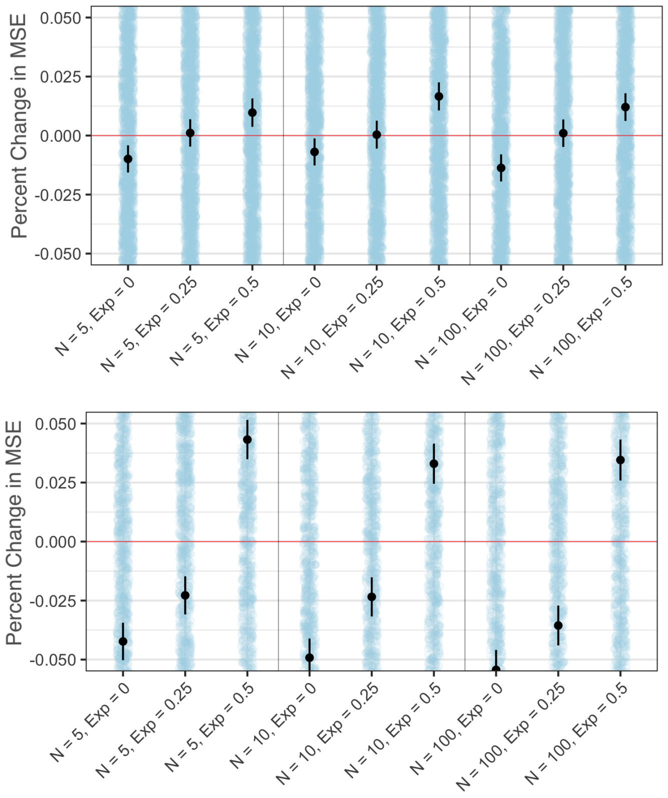

Fig. 3. Simulations assessing the effect of group size and expert responsivity on group accuracy in the presence of peer communication. Each group has N individuals, and agents update according to Equation 3. Exp determines the amount of variation in σ_1^2 and γ, where $Exp = 0$ means there is no variation in responsivity with expertise (all σ_1^2 and γ are equal), while $Exp = 0.5$ means there is high variation (some agents with high expertise, ie low σ_1^2 and high γ, and others with the reverse). See SI for details. Top: Change in MSE, N exchanges, as percentage of average MSE (ie, $(MSE_{pre} - MSE_{post})/(MSE_{pre} + MSE_{post})/2$. Bottom: 10N exchanges.

fact. The most common approach is to reweight predictions based on various observed features of the individuals, such as demographics or treatment conditions (Mellers et al., 2015; Moore et al., 2016; Kelley and Tetlock, 2013; Mannes et al., 2014; Budescu and Chen, 2015). We tested the effects on collective accuracy in two ways, first by reweighting predictions by each individual feature separately, and then by reweighting using all features at once using machine learning methods. In both cases, we found that the best weighting can increase collective accuracy by about 10%, which is substantial, although less than what is possible by simply increasing N (since the collective error, like the standard error, decreases with $\sqrt{N}$).

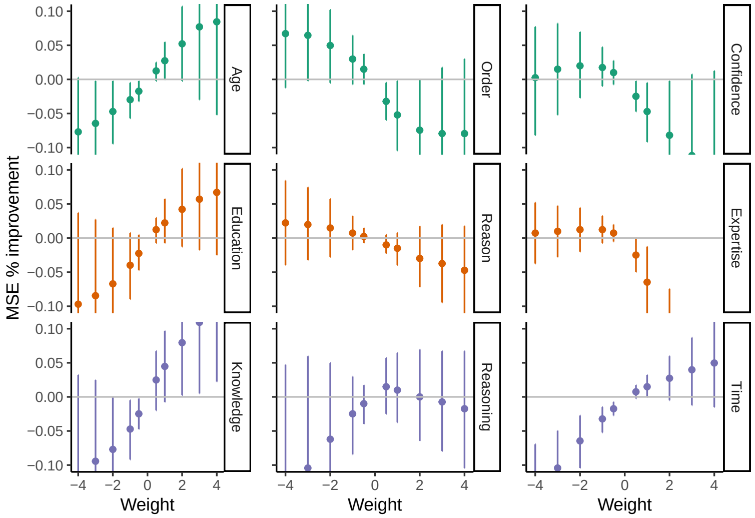

Fig. 4. The effects on aggregate accuracy of weighting predictions by individual features. Each weight w means that that (standardized) feature f was given weight f^w with weighted prediction $p_i' = p_i * f_i^w$. Higher is better, and errors are bootstrapped at the individual level.

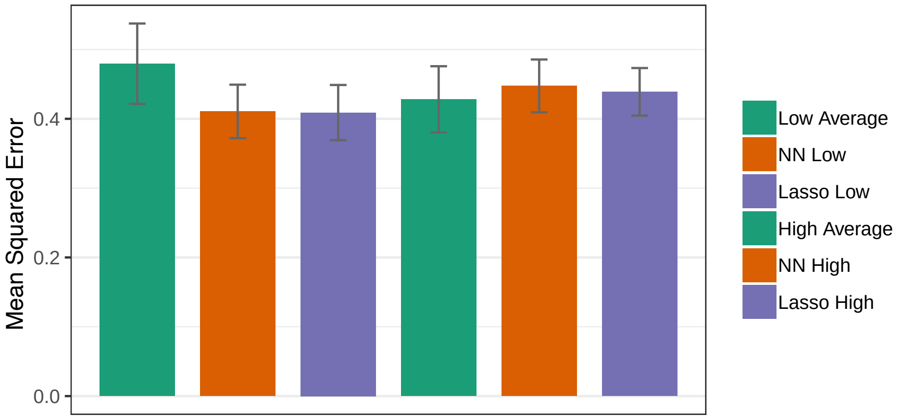

Fig. 5. Out-of-sample prediction accuracy using a two-layer deep neural net and lasso regression to reweight subject predictions by all measured features. Columns 1 and 4: unweighted; columns 2 and 5: Low- and high-expertise subjects reweighted by the neural network; columns 3 and 6: Low- and high-expertise subjects reweighted by the lasso regression. Lower scores are better.

As Figure 4 shows, taken individually, most effective was to assign more weight to subjects with higher age, education, and political knowledge; also effective was to assign less weight to questions answered later in the (random) sequence, presumably due to subject fatigue with later questions. To weight all features jointly, we fed predictions, all covariates, and interactions between the two into a two-layer neural network or a lasso regression, both of which allow predictions to interact with, and thus be effectively reweighted by, all individual and task-level covariates. The model was trained to maximize out-of-sample group accuracy, where the omitted samples were at the question level (ie, the model was fit on data from responses to 8 questions and tested on 8 held-out questions, with repeated question-level resampling to generate bootstrapped errors). As can be seen in Figure 5, before reweighting, the expert group has about a 10% lower out-of-sample error than the non-expert group (column 1 vs 4). However, after reweighting, the experts are unimproved, but the non-experts become as accurate as the experts (columns 2-3 vs 5-6). So whether reweighting by individual features or across all features, the improvement is perhaps 10%. This benefit, however, accrues mainly to the non-experts, raising their collective accuracy to a level comparable to experts – albeit "experts" only insofar as they are in the upper half of the skill distribution.

However, our Bayesian framework also suggests another method for improving collective accuracy. Since treated subjects are strongly affected by the anchor value they saw, and since as our simulation shows, seeing a peer prediction can in many cases be detrimental to collective accuracy, we therefore hypothesized that peer-affected predictions could in a sense be "corrected" back towards the original, unaffected prediction. This is of course only possible because we know the treatment values, but it does not require us to know the pre-treated predictions: if an individual's post-treatment position is $a' = (1-\gamma)a + \gamma b$, then their original, "corrected" position $a = (a' - \gamma b)/(1-\gamma)$. Figure 6 shows the results on group accuracy of de-biasing all treated predictions using γ values ranging from 0 to 1, with bootstrapped errors. Pooling all treated subjects, the boost to collective accuracy due to this correction again peaks at around 10%, at a $\gamma = 0.4$, which is not far from our measured anchor effect of 0.3. Dichotomizing by our usual measure of expertise (accuracy in the first half of questions answered) leads to very large errors since it requires throwing out half of all responses per subject in addition to the 50% of subjects who were not treated with a peer prediction, but dichotomizing by our other expertise measure, reasoning skill (accuracy on the three-question test), shows that the more expert group is at best only mildly improved, while the less expert group derives all the benefit. These differences in improvement are presumably because the experts are less affected by peer information, and possibly somewhat benefited by internal differences in responsivity (akin to the $Exp = 0.25$ regime in Figure 3), while the non-experts and group as a whole are more harmed by peer information (akin to the $Exp = 0.0$ regime in Figure 3), and thus have more room for improvement.

Note that because our treated subjects received anchors from both previously treated subjects, and untreated subjects, the effective pool of information for the anchored group is higher than the unanchored group. This did not increase the collective accuracy of the anchored group, and if anything, it should make it harder for the correction procedure to improve collective accuracy, since in effect the correction does away with treatment information in restoring pre-treatment predictions. However, the harms due to intra-group communication appear to outweigh the

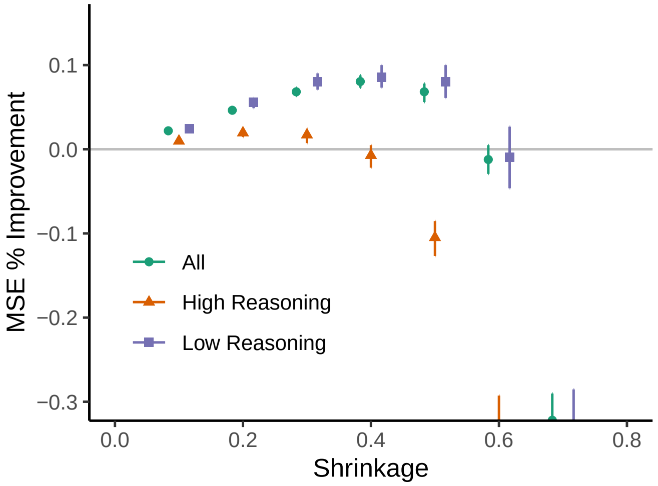

Fig. 6. Accuracy improvement from “correcting” predictions by shrinking each towards the question mean by the specified percentage (95% confidence intervals from bootstrapped standard errors).

benefits due to extra-group information plus the benefits due to implicit expert weighting, so our results suggest that in purely intra-group communication, the correction benefits may be even higher. Our correction increases group accuracy in part by re-inflating group variance, which is diminished by intra-group communication: eg, for $\gamma = 0.5$, variance is increased from a post-treatment value of 0.5 up to 2.0, vs 0.8 for the unanchored group. We hypothesize that this re-inflation may be more effective and less likely to overshoot when applied to a purely intra-communicating group, but characterizing analytically precisely what debiasing amount will work best given subject expertise, belief distributions, persuasive effects, and expert responsivity, is an important question for future work.

Discussion

The reduced responsivity of experts to peer information and greater selectivity towards information quality appear to be general properties of expertise that hold for textual as well as numerical signals. Numerically, our Bayesian framework provides a unified model in which experts discount information according to two separate processes: First, a generally increased discounting of all information due to a higher confidence in their own beliefs; and second, a tendency to more steeply discount more distant information. While previous work has demonstrated both of these results in broad strokes, our real-world, continuous-valued prediction task allows us to precisely estimate a continuous response function, and demonstrate its qualitative similarity to our theoretical Bayesian framework.

Our results also show for the first time that these discounting tendencies hold not just for numeric communication, but also for linguistic communication: experts are less sensitive to textual information overall, and also more selective towards higher-quality information. Previous numeric work has necessarily left it somewhat ambiguous whether the tendency of experts to more steeply discount discordant information is due to a preference for higher-quality, or simply to a tendency to discount any information in proportion to its distance from the expert’s prior. Our textual results suggest that it is indeed quality that experts are sensitive to.

Our experimental design also allows us to trace the effects of peer information, and the mediating effect of expertise, through to individual and collective accuracy. While one might expect that less-expert subjects would be most benefited by peer communication, it is also possible that the reduced sensitivity and increased selectivity of experts to peer information could mean that more-expert subjects may in fact benefit more. We find that these individual benefits very with both expertise and information quality: while the less-expert are indeed benefited more by numeric and textual peer information overall, experts are in fact benefited by the highest quality textual information. Expert selectivity means that they do benefit from communication even when much of that information may be relatively poor. Our result showing that experts benefit via selectively responding to high-quality text is particularly important, since with purely numerical information, it is difficult for discounting alone to provide individual benefits for experts.

Finally, when turning to aggregate, "wisdom of the crowd" accuracy, our Bayesian framework allows us to show both theoretically and via simulation that the collective benefits are highly sensitive to group size, number of exchanges, prior distributions, and expert responsivity. In our experimental setting, aggregate accuracy was neither helped nor harmed by receiving peer communication. Our simulations suggest that, in addition to being due to a small number of groups to test (16), this may also be due to the limited number of exchanges and relatively low difference in expert responsivity in our experiment. Pragmatically, it suggests that the implicit weighting of experts due to their reduced responsivity may be boosted by increasing the number of exchanges, decreasing expert responsivity via enhancing successful reputations, and direct interventions to reduce non-expert to expert communication in cases where expertise can be independently measured.

Like others (Mannes et al., 2014; Budescu and Chen, 2015; Mellers et al., 2015; Madirolas and de Polavieja, 2015; Budescu and Chen, 2015; Tetlock and Gardner, 2016; Luo et al., 2018), we also found that group accuracy could be improved by reweighting individuals, either by individual features, or jointly across all features using machine learning. This reweighting appears to only help the less-expert group, raising its collective accuracy to the level of the expert group. However, our experimental design and Bayesian framework also suggest a novel form of collective improvement, by reconstructing each individual's (unseen) original prediction from their post-treatment prediction and observed signal. The benefits of this procedure have significant limitations though: First, only the accuracy of the less-expert group could be meaningfully improved, presumably due to their higher overall responsivity. Second, this procedure only benefits intra-group communication: if most signals are coming from outside the group, the harms of destroying that information likely outweigh the benefits of restoring unbiased predictions. And third, this procedure can only work when one has completely tracked all signals exchanged by participants. The third issue may not be as much of a problem in the modern era though, where a prediction platform can readily track every page a user sees. In such areas, the primary benefit of this correction procedure would be to allow the social benefits of communication – benefits without which almost no real-world collaborative platform can long survive – while at the same time, restoring much of collective accuracy lost due to that communication.

As discussed earlier, in order to achieve the benefits of reweighting or prediction correction, it is necessary to have an accurate, domain-specific measure of expertise, not just a measure of general knowledge or education (and self-assessment is if anything anticorrelated with true expertise). But perhaps surprisingly, even in our experimental framework where the average subject answered only 8 questions, and therefore the average number of questions used to calculate "expertise" was only 4, this rough measure is sufficiently strong to demonstrate the myriad mediating effects of expertise. The effectiveness of this measure is presumably in part due to our relatively large subject pool, which allows us to identify small but consistent effects, but it may also be due to our continuous-valued prediction tasks, which allow for far more variation per subject than would, eg, using an aggregate of four binary correctness scores.

The relative ease of measuring expertise, at least roughly, implies that there may be many practical opportunities to construct communication structures that maximize collective accuracy, such as those which only allow information flow from experts to non-experts, even if specific individuals participate only briefly in the platform. Similarly, as Figure 2 shows, non-experts are hurt by being asked to provide a reason ("Reason Requested"), while experts are not, which suggests that an ideal textual communication system may similarly entail structuring the flow of information from experts (who are not harmed by requesting reasons) to non-experts (who are harmed by requests but benefit most from receiving information).

The weaker but similar mediating effects we found with our simple 3-question "reasoning skill" measure also suggests that there may be scope for developing better general-purpose measures of expertise, at least with regards to forecasting. This is an important target for future work. But when expertise is not readily measurable, our results suggest that the benefits of the natural upweighting of experts due to their reduced responsivity can be maximized by increasing the number of exchanges and by boosting the variance in expert responsivity, such as by diversifying the subject pool or enhancing reputations. These structural changes may provide many of the benefits of expert weighting and hierarchical information flow without the need to directly measure expertise, and may allow less-expert groups to enhance their performance nearly to the level of experts.

Bibliography

Atir, S., Rosenzweig, E., and Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, 26(8):1295–1303.

Attali, Y., Budescu, D., and Arieli-Attali, M. (2020). An item response approach to calibration of confidence judgments. *Decision*, 7(1):1.

Becker, J., Brackbill, D., and Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076.

Becker, J., Porter, E., and Centola, D. (2019). The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22):10717–10722.

Budescu, D. V. and Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280.

Cheek, N. N. and Norem, J. K. (2017). Holistic thinkers anchor less: Exploring the roles of self-construal and thinking styles in anchoring susceptibility. *Personality and Individual Differences*, 115:174–176.

Dunning, D., Johnson, K., Ehrlinger, J., and Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current directions in psychological science*, 12(3):83–87.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

Golub, B. and Jackson, M. O. (2010). Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49.

Guilbeault, D., Becker, J., and Centola, D. (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences*, 115(39):9714–9719.

Guilbeault, D. and Centola, D. (2020). Networked collective intelligence improves dissemination of scientific information regarding smoking risks. *PLOS ONE*, 15(2):1–14.

Hong, L. and 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(46):16385–16389.

Jayles, B., Kim, H.-r., Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., Sire, C., and Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, 114(47):12620–12625.

Kao, A. B., Berdahl, A. M., Hartnett, A. T., Lutz, M. J., Bak-Coleman, J. B., Ioannou, C. C., Giam, X., and Couzin, I. D. (2018). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface*, 15(141):20180130.

Kaustia, M., Alho, E., and Puttonen, V. (2008). How much does expertise reduce behavioral biases? the case of anchoring effects in stock return estimates. *Financial Management*, 37(3):391–412.

Kelley, E. K. and Tetlock, P. C. (2013). How wise are crowds? insights from retail orders and stock returns. *The Journal of Finance*, 68(3):1229–1265.

King, A. J., Cheng, L., Starke, S. D., and Myatt, J. P. (2011). Is the true 'wisdom of the crowd'to copy successful individuals? *Biology Letters*, 8(2):197–200.

King, A. J., Cheng, L., Starke, S. D., and Myatt, J. P. (2012). Is the true 'wisdom of the crowd'to copy successful individuals? *Biology Letters*, 8(2):197–200.

King, A. J. and Cowlishaw, G. (2007a). When to use social information: the advantage of large group size in individual decision making. *Biology letters*, 3(2):137–139.

King, A. J. and Cowlishaw, G. (2007b). When to use social information: the advantage of large group size in individual decision making. *Biology Letters*, 3(2):137–139.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Luo, Y., Iyengar, G., and Venkatasubramanian, V. (2018). Social influence makes self-interested crowds smarter: An optimal control perspective. *IEEE Transactions on Computational Social Systems*, 5(1):200–209.

MacCallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, 7(1):19.

Madirolas, G. and de Polavieja, G. G. (2015). Improving collective estimations using resistance to social influence. *PLOS Computational Biology*, 11(11):1–16.

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276.

Mavrodiev, P., Tessone, C. J., and Schweitzer, F. (2013). Quantifying the effects of social influence. *Scientific Reports*, 3(1):1360.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., and Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, 21(1):1.

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

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H., and Tenney, E. R. (2016). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, 63(11):3552–3565.

Mussweiler, T. and Strack, F. (2000). Numeric judgments under uncertainty: The role of knowledge in anchoring. *Journal of experimental social psychology*, 36(5):495–518.

Nobre, D. A. and Fontanari, J. F. (2020). Prediction diversity and selective attention in the wisdom of crowds. *Complex Systems*, 29(4):861–875.

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., and Wirth, K. (2016). Random number simulations reveal how random noise affects the measurements and graphical portrayals of self-assessed competency. *Numeracy: Advancing Education in Quantitative Literacy*, 9(1).

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., and Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy: Advancing Education in Quantitative Literacy*, 10(1).

Page, S. E. (2018). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton University Press.

Patry, M. W. (2008). Attractive but guilty: Deliberation and the physical attractiveness bias. *Psychological reports*, 102(3):727–733.

Schlösser, T., Dunning, D., Johnson, K. L., and Kruger, J. (2013). How unaware are the unskilled? empirical tests of the "signal extraction" counterexplanation for the dunning–kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, 39:85–100.

Sommers, S. R. (2006). On racial diversity and group decision making: identifying multiple effects of racial composition on jury deliberations. *Journal of personality and social psychology*, 90(4):597.

Sunstein, C. R. and Hastie, R. (2015). *Wiser: Getting beyond groupthink to make groups smarter.* Harvard Business Press.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business*. Doubleday New York.

Tetlock, P. E. and Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Toyokawa, W., Whalen, A., and Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, 3(2):183–193.

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

Welsh, M. B., Delfabbro, P. H., Burns, N. R., and Begg, S. H. (2014). Individual differences in anchoring: Traits and experience. *Learning and Individual Differences*, 29:131–140.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1):1 – 13.