The contingent wisdom of dyads:

When discussion enhances versus undermines the accuracy of collaborative judgments

Julia Minson

Harvard University

Jennifer Mueller

University of San Diego

Richard Larrick

Duke University

Published in **Management Science in 2017

ABSTRACT

We evaluate the effect of discussion on the accuracy of collaborative judgments. In contrast to prior research, we show that discussion can either aid or impede accuracy relative to simple averaging of collaborators' independent judgments, as a systematic function of task type and interaction process. On estimation tasks with a wide range of potential estimates, discussion aided accuracy by helping participants prevent and eliminate egregious errors. On estimation tasks with a naturally bounded range, discussion following independent estimates performed on par with averaging. Importantly, if participants did not first make independent estimates, discussion significantly harmed accuracy by limiting the range of considered estimates, independent of task type. Our research shows that discussion can be a powerful tool for error reduction, but only when structured appropriately: Decision-makers should form independent judgments in order to consider a wide range of possible answers, and then use discussion to eliminate extremely large errors. (149 words)

The contingent wisdom of dyads:

When discussion enhances versus undermines the accuracy of collaborative judgments

A large body of research in judgment and decision-making rests upon a premise that is elegant in its simplicity: namely, that the combined estimates of multiple individuals are more accurate than those made alone (Ariely et al., 2000; Clemen, 1989; Hogarth, 1978; Larrick & Soll, 2006; Lorge, Fox, Davitz, & Brenner, 1958; Makridakis & Winkler, 1983; Simmons, Nelson, Galak, & Frederick, 2011; Surowiecki, 2005). This notion that "two heads are better than one" frequently guides how corporations, groups, and families go about making decisions. An open debate, however, exists around whether discussion, arguably the most common decision-making process outside of the psychology laboratory, aids or harms judgment accuracy (Minson & Mueller, 2013; Schultze, Mojzisch, & Schulz-Hardt, 2013).

Although prior work on this topic has spanned the gamut of group sizes from those comprised of three individuals to markets comprised of thousands, we focus on the smallest collaborative unit: the dyad. Dyadic judgment is of both theoretical and practical interest. Specifically, the standard error of the average of several judgments decreases proportionally to the square root of the number of averaged judgments. Thus, the transition from individual to dyadic judgment represents the largest possible increase in accuracy that can result from the addition of a single person. Simply put, doubling of the "sample size" of judges dramatically reduces error. Psychologically, dyads are also unique because they cannot use most strategies available to larger groups in order to decide how to weight members' estimates. Due to the absence of a "majority" each judgment in a dyad must be weighted on its own merits, and not as a function of how well it conforms to the judgments of other group members.

Furthermore, dyadic judgment is extremely common. From the dyadic organization of most adult households to the well-publicized successes and failures of Larry Page and Sergey Brin (Google), Ben Cohen and Jerry Greenfield (Ben and Jerry's), Bill Gates and Paul Allen (Microsoft), John Reed and Sandy Weill (Citigroup), and Steve Jobs and Steve Wozniak (Apple) millions of decisions every day are made by groups of two.

To the extent that dyadic discussion is common in judgment and decision-making contexts, is it beneficial to judgment accuracy? The creativity literature notes that in the context of "idea pitches" – "pitchers" – those who pitch ideas, and "catchers" those who vet ideas for funding - often engage in collaborative discussion which aids catchers' ability to diagnose creative potential (Elsbach & Kramer, 2003). This perspective is further supported by a large body of research on groups and teams suggesting that discussion is the vehicle through which collaborators integrate divergent viewpoints to generate new ideas (Ancona & Caldwell, 1992; Beersma & De Dreu, 2005; De Dreu, 2006; De Dreu & West, 2001; Kurtzberg, 2005) and high quality decisions (De Dreu, Nijstad, & van Knippenberg, 2008).

However, empirical research in judgment and decision-making tends to reach a different conclusion. Most frequently, studies evaluating group discussion have found the accuracy of discussed judgments to be equivalent to the accuracy of the simple average of the earlier independent estimates of the same individuals (see Gigone & Hastie, 1997, for a review). Partly based on this evidence, researchers have suggested that the real effect of discussion may be to simply boost confidence in one's previous positions (Heath & Gonzales, 1995) and that face-to-face meetings should be eliminated entirely (Armstrong, 2006). Recently, the insight that averaging [1]multiple individual judgments is more effective than most forms of deliberation has even received substantial attention in the popular scientific and management press (e.g., Sunstein, 2006; Surowiecki, 2005).

Although there are many differences between the creativity and the judgment literatures, upon closer examination the long-standing conclusion that discussion does not outperform the averaging of independent estimates is based on a body of work that is not entirely consistent. In their seminal review of this research, Gigone and Hastie included seventeen empirical papers and re-analyzed their own data. As the authors point out, most of the published papers do not actually include sufficient information to compare the accuracy of group judgments relative to the averaging benchmark. Instead, many papers compare the accuracy of group judgments to the average accuracy of individual judgments. In addition to

[1] In the present manuscript we use the term "average" to refer to the arithmetic mean of multiple estimates.

not being the appropriate comparison, this also introduces terminological confusion. The comparison most commonly reported is based on calculating the errors of individuals and then averaging those errors in order to establish a performance benchmark for a typical individual working alone. Of course as Gigone and Hastie point out, the appropriate comparison against which discussed estimates *should* be judged is the error of the averaged judgments of all group members.

Of the papers that do make this correct comparison, only one article testing an elaborate group process intervention reports a set of group judgments that outperform the average of independent estimates (Reagan-Cirincione, 1994). Based on these findings and a detailed reanalysis of their own relevant data Gigone and Hastie conclude that simply averaging independent judgments of group members produces a level of accuracy that is generally indistinguishable from what the same group members can achieve through discussion.

However, the literature *also* provides us with several studies demonstrating the opposite pattern of results. Specifically, discussed group judgments *do* outperform simple averaging in Sniezek & Henry (1989), Minson, Liberman & Ross (2011; studies 1 & 2), and Schultze, Mojzisch & Schulz-Hardt (2012). Yet, although these three papers describe results that lead to opposite conclusions than those of Gigone and Hastie, this discrepancy has gone largely unnoticed and un-reconciled.

In the present research we explore the basic question of whether discussion helps or harms judgment accuracy and demonstrate that these effects are more complex than either lay intuition or the empirical evidence suggest, varying systematically with the nature of the task and the structure of the estimation process. In contrast to most prior empirical work, Study 1 shows that discussion can reliably outperform the accuracy of averaging for particular estimation tasks. Specifically, discussion outperforms averaging on tasks in which individuals working alone have a high likelihood of committing an egregious error as might occur when judges have little domain-specific expertise or are operating under conditions of very high uncertainty. This potential for extreme errors is an important task feature that moderates the benefits of discussion.

The relevance of this factor emerges if one compares previous studies. For example, studies by Gigone and Hastie that examined the effect of discussion used a familiar domain on a limited scale (student grades), allowing for limited errors. By contrast, Sniezek & Henry (1989) asked groups to estimate the frequency of particular causes of death in the United States based on a total population of 230,000,000. Given that the target quantities included high frequency causes of death such as heart disease and lung cancer, as well as rare causes such as smallpox, participants had ample opportunity to make very large errors. As our data demonstrate, the frequency of such large errors accurately predicts the relative effectiveness of averaging versus discussion. We go on to show that this occurs because large errors both decrease the benefits of averaging and are easily eliminated in the course of discussion.

Study 2 further explores the mechanism behind this result, while also identifying a moderator that can make discussion *underperform* averaging. Specifically, we manipulate whether participants engage in discussion following independent estimates (as has been the case in previous research) or prior to making independent estimates (as often happens outside of the laboratory). Logically, if discussion can be used to eliminate large individual errors, it may serve individuals even better if it is used to prevent such errors. Indeed, lay people tend to believe that the best process for generating accurate judgments is to discuss those judgments with a partner and then discuss them again. However, a large literature on anchoring and judgment assimilation suggests that if discussion takes place prior to independent estimation, it will hurt accuracy by limiting the range of values considered. In Study 2 we find that discussion that precedes independent judgments does in fact impair dyadic performance.

The Benefits and Challenges of Combining Judgments

Early research on judgment aggregation statistically combined the estimates of individual participants with no opportunity for discussion or interaction of any kind (Galton, 1907; Klugman, 1947; Knight, 1921). And although this work has generated a recent resurgence of interest (Surowiecki, 2005), the key insight from the 1930's holds still: Combining multiple judgments improves accuracy because uncorrelated errors associated with individual judgments cancel out (Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Vul & Pashler, 2008). This tendency, sometimes called the "wisdom of crowds effect"

(Lorenz et al., 2011; Surowiecki, 2005), makes simple averaging a powerful and efficient strategy for error reduction.

However, it is possible that having access to the judgments of multiple individuals could further reduce error beyond the level available through simple averaging, if individuals were able to weight collaborators' inputs based on their relative accuracy (Minson, Liberman, & Ross, 2011). This, in turn, would require a consistent ability to distinguish more versus less accurate judgments, even under conditions of uncertainty. Some ability to give greater weight to more accurate judgments has been documented (Minson et al., 2011; Minson & Mueller, 2012). If conditions were right to enable individuals to sift through available cues and sort them according to quality, discussion could become a powerful tool for producing judgments that outperform simple averaging.

In the present research we hypothesize that an important determinant of the effectiveness of averaging versus selective weighting via discussion is the *range* of estimates considered by participants. On one hand, when the initial estimates entertained by collaborators cover a wide range (because of dispersed knowledge, or high uncertainty, or low expertise), the truth is more likely to lie within that range, making averaging highly effective. However, a wide range also means that at least some of the collaborators' estimates contain large errors. And whereas such errors may cancel when the "crowd" is sufficiently numerous, they are unlikely to do so in the types of small groups that comprise most management teams, entrepreneurial partnerships, or households. When large errors are present and the pool of judgments is small, the effectiveness of averaging will be limited and the ability to weigh inputs judiciously will gain importance.

Our studies provide evidence that under these conditions discussion can be effective for weighting inputs of disparate quality. Discussion enables collaborators to explain the bases for their estimates, share their level of confidence, and question each other's assumptions, as well as catch mistakes that arise from inattention or misunderstanding. And although prior research has generally found that error reduction through discussion does not outperform averaging, we demonstrate that the value of discussion increases dramatically in contexts where large errors are common.

Research Overview

In two studies we compare the judgment accuracy resulting from discussion to that resulting from averaging. In Study 1 we used three different participant samples making estimates in different domains to demonstrate that discussion can reliably outperform averaging when large errors are common. We are further able to show that the reduction in error that discussing dyads accomplish relative to averaging is specifically due to elimination of egregious errors. Additionally, we use a resampling simulation to compare the discussed estimates of dyads to averaged estimates of various sized groups to show how discussion performs relative to averaging a larger sample of estimates.

In Study 2 we experimentally manipulated participants' likelihood of committing an extremely large error in order to establish the causal role of the error distribution in the benefits of discussion. We also tested a second moderator of the benefits of discussion on judgment accuracy. Namely, we manipulated whether discussion occurs after each dyad member forms an independent estimate (as in previous studies), or in the absence of making an independent estimate (as often happens outside of the laboratory). We find that discussion in absence of independent estimates is actually *detrimental* to accuracy. We go beyond prior work by analyzing recorded discussions to document underlying cognitive and social processes. In our concluding discussion we also consider alternative methods of aggregating estimates including medians of larger groups and geometric means.

Study 1

Study 1 tested whether the accuracy of discussed estimates outperforms that of averaging in contexts in which there is a high probability of individuals committing an extremely large error. Although it is an open question how often such contexts arise, recent work by Lobo and Yao (2010) using seventeen large data sets demonstrates that human judgments often do not reflect a Gaussian Normal distribution, but are "heavy-tailed," with frequent and large extreme errors. Indeed, fields such as economic forecasting, marketing, and many of the natural sciences are rife with estimates and predictions that must be made under conditions of extreme uncertainty. Thus, if discussion leads to improved

accuracy in cases where very large errors are common, the role of discussion should be re-evaluated more generally since such cases might be the norm rather than the exception.

We used three different participant samples and three different estimation domains to test the reliability of our results. Our dataset also allows us to explore the extent to which the benefits of discussion are particularly large in cases when one of the collaborators committed an extreme error. Furthermore we use a resampling simulation to compare the effectiveness of dyadic discussion to averaging larger samples of participant judgments.

Method

Participants. Participants (N = 180) were members of a university research pool compensated with $7. Sample A estimated the annual salaries of nine Fortune 500 CEOs (n = 68). Sample B estimated weights of nine animals based on their photographs (n = 52). Sample C estimated nine quantities related to US demographic statistics (n = 60), (Table 1). Participants made estimates for all items individually (Round 1), and then worked with a randomly-assigned partner to make a new set of joint estimates through discussion (Round 2). This procedure enabled us to compare the accuracy of individual estimates to that of estimates made jointly through discussion, and to that of the average of the earlier individual estimates.

Variables. To establish a metric that would allow us to compare accuracy across estimation items on different scales, we first calculated the absolute difference of each individual estimate from the correct answer for that item. Smaller absolute differences are more accurate. We then used the mean and standard deviation of those absolute differences within each item to standardize the accuracy of averaging and of the estimates reached through discussion. Although this makes it more difficult to appreciate the absolute magnitude of participant errors on any given item, our specific interest lies in the *relative* levels of accuracy of discussed versus averaged estimates. Thus standardizing each error relative to the typical error for that item ensures that items with particularly large or small absolute errors do not skew our results.

In order to calculate the frequency of extreme errors we noted the number of estimates that were an order of magnitude (i.e. ten times greater or ten times smaller than) away from the truth.

For each item, participants reported their confidence that their individual or their jointly-produced estimate fell within 10 percentage points of the answer on a five-point scale from *"Not at all confident"* to *"Extremely confident."* We audiotaped all discussion during the estimation task.

Analytical approach. In this study and in Study 2, we clustered our analyses at the dyad level because each participant provided nine estimates and worked with a partner. In both studies, we entered item fixed effects to control for item-specific variance. We report bootstrapped standard errors based on 2000 repetitions (Efron & Tibshirani, 1994; Mooney, Duval, & Duvall, 1993) because our data are not normally distributed. Our target sample size was 30 dyads per sample. Variation from that number is due to participants signing up (or not) during the scheduled time. Eight pairs of estimates were excluded from analysis because dyad members did not enter the same estimates in Round 2 (when they were instructed to reach agreement). In both studies, we analyzed all complete observations, and report all manipulations and measures (Simmons, Nelson, & Simonsohn, 2012).

Results

In each of our samples estimation accuracy improved from Round 1 individual estimates to Round 2 joint estimates. Figure 1 shows the reduction in error possible through averaging and that which resulted from discussion, scaled relative to the errors of individual estimates as described in the Variables section above. The error of Round 2 estimates was consistently lower than the error of estimates that could have been reached through averaging. This difference was observed in all of our data combined, and in each sample separately (Combined data: $b = 0.10$, $z = 5.07$, $p < .001$, [0.06, 0.14]; Sample A: $b = 0.13$, $z = 3.22$, $p = .001$, [0.05, 0.21]; Sample B: $b = 0.08$, $z = 3.16$, $p = .002$, [0.03, 0.13]; Sample C: $b = 0.10$, $z = 2.80$, $p = .005$, [0.03, 0.17]).[2]

[2] Whereas our graphs plot reductions in error, where a negative difference represents improvement, in our analyses of round to round accuracy change we code accuracy improvement as positive.

To begin exploring the role of extreme errors in the relative benefits of discussion versus averaging, we regressed the amount of improvement that dyads achieved in the course of discussion relative to simple averaging on a binary variable representing whether at least one of the dyad members' initial estimates constituted an extreme error (i.e. was ten time larger or ten times smaller than the correct answer). Across our three samples, we observed a positive relationship between whether at least one of the estimates in the dyad constituted an extreme error and the extent to which discussion outperformed averaging (b = .21, z = 3.34, p = .001, [0.09, 0.34]).[3] In fact, an examination of all judgments at the dyad level shows that the improvement attained from discussion was almost entirely attributable to the instances in which an individual estimate contained a large error (n = 630, mean improvement = .25), with almost no improvement observed in the instances in which there was no large error (n = 990, mean improvement = .01).

Although our data suggest that discussion can indeed outperform averaging, there remains the critical question of whether this accuracy benefit is attributable to the effectiveness of discussion or to the *in*effectiveness of averaging in the presence of extreme errors. To address this we specifically examined only the cases in which participants committed errors that fell on the same side of the truth, i.e. did not "bracket" the answer (Larrick & Soll, 2006). In such non-bracketing cases, the error of the average of two estimates has to be exactly equal to the average error of the two individual estimates. Therefore, any systematic improvement of discussion relative to averaging has to be due to the fact that dyad members gave greater weight to better information. Indeed, when we compare the accuracy of averaging to the accuracy of estimates produced through discussion in these cases (n = 1,104 estimates), we see that across our three samples, discussion continued to be more accurate than averaging (b = 0.12, z = 5.91, p < .001, [0.08, 0.16]).

A final question concerns whether averaging the estimates of a larger group of participants, rather than only two, would have produced results that compared favorably to discussion. To address this we

[3] We obtain similar results if we define extreme errors differently, i.e. as seven times larger or smaller than the truth (or 20 times larger or smaller).

randomly sampled with replacement 10,000 groups of estimates from our data in order to simulate the averaged estimates of groups of 3-10 participants. On 7 of the 27 items the averaged estimates of groups of 3-10 participants would have performed better than discussion. Only on 2 items did averaging the estimates of two individuals outperform discussion. On the remaining 18 items used in Study 1, no group of any size, including all 50+ participants, performed better than dyadic discussion. Because these estimation topics produced large errors, averaging did not perform well, even in fairly large groups. However, a brief discussion between two people was often sufficient to reduce these large errors and confer substantial accuracy benefits.

Discussion

Study 1 demonstrated that across three different estimation domains attempted by three different samples of participants, discussion outperformed simple averaging. Further analyses demonstrated that this reduction in error was driven almost entirely by cases where one or both of the collaborators made an extremely large error when working individually, and that these errors were eliminated in the course of a brief discussion.

One might imagine that the reason discussion beats averaging when extreme errors are present is because averaged estimates that are based on highly erroneous individual estimates are just not very accurate. However, our data demonstrate that even in cases where both individual estimates err on one side of the truth discussion leads to improvement. Because in such cases the average error of the two estimates is exactly equal to the error of averaging, improvement relative to averaging *has* to mean that participants are giving greater weight to more accurate input. This is strong evidence for the fact that our effect is due at least in part to the effectiveness of discussion, not merely to the ineffectiveness of averaging.

Our Study 1 results raise two important questions. First, since we relied on endogenous variation in the occurrence of extreme errors, we cannot conclude the presence of a causal relationship between extreme errors and the value of discussion. We address this issue in Study 2 by manipulating the scale of our estimation items in order to elicit a lower or higher rate of extreme errors.

Second, our estimation process in Study 1 follows procedures regularly used in prior studies wherein participants made independent estimates and then used discussion to revise those estimates. This process may not be particularly common in the world outside of psychology laboratory if individuals hold lay theories about the value of "keeping an open mind," or the limited value of individual contemplation.

Conceptually, it may be the case that discussion that occurs prior to individual judgments can serve to reduce the rate of individual extreme errors. For example, discussion could help decision makers identify relevant facts, valid cues, and useful examples in a decision problem. If the problem entails estimating the number of people without health insurance in a major city, a general knowledge of city sizes and the proportion of the population that is uninsured both help bound the problem. If discussion prevents the occurrence of large errors, than the question of averaging independent judgments versus combining them through discussion is moot – individuals should simply start out with discussion. Thus in Study 2 we manipulate the timing of discussion relative to independent estimation to establish which process is more effective.

Study 2

Because Study 1 suggests that discussion outperforms averaging in contexts where individual estimation may produce very large errors we manipulated the probability of large errors in Study 2 by varying the phrasing of estimation items. Participants made estimates regarding a particular topic (e.g. the insurance status of Philadelphians) on an unbounded scale that allowed for the commission of extreme errors ("How many Philadelphians have no health insurance?"), or on a scale that was bounded and thus limited their ability to make such errors ("What percentage of Philadelphians has no health insurance?").

We also manipulated the estimation process by requiring participants to make independent estimates before discussion (following prior research), or by instructing them to begin discussion without prior independent estimates. This latter process was favored by online respondents in a set of pilot studies we conducted and is one that we suspect to be common outside of the laboratory.

The benefits of discussion relative to averaging documented in Study 1 can be thought of as arising from the effectiveness of discussion or the *in*effectiveness of averaging estimates containing large

errors. It may be the case that discussion preceding individual estimates can prevent participants from committing such errors due to more effective problem framing and assumption testing (De Dreu et al., 2008). This may lead to greater accuracy and time savings relative to a discussion that starts with error-ridden individual estimates.

However, several research streams suggest that discussion without independent judgments may in itself be costly because it may limit the range of estimates considered by collaborators. Under uncertainty, individuals succumb to both normative and informational social influence (Deutsch & Gerard, 1955). To the extent that collaborators use anchoring and insufficient adjustment as a heuristic (Koehler & Beauregard, 2006; Tversky & Kahneman, 1974), they are likely to be swayed by the first estimate spoken (Koehler & Beauregard, 2006). Indeed, recent research demonstrates that simple exposure to others' estimates diminishes the wisdom of crowds because it leads to correlated error (Lorenz et al., 2011). Thus, if discussion leads to assimilation, it may limit the accuracy benefits of collaboration by decreasing the likelihood that the considered range is large enough to bracket the truth (cf., Minson, & Mueller, 2012; Schultz et al 2012).

Given the above possibilities, in Study 2 we orthogonally manipulated the likelihood that participants would make an extreme error, and had the opportunity to make individual estimates prior to engaging in discussion. The resulting factorial design allows us to examine how the nature of the estimation task and the manner in which the judgment process is structured jointly determine discussion outcomes. Additionally, in Study 2 we incentivized participants for accuracy in order to ensure that they put careful thought into their estimates.

Method

Participants. Participants (N = 228) were members of a university research pool, compensated with $10. To incentivize accuracy, we offered a $30 bonus on each of the two estimation rounds, which decreased by $1 for each percentage point any estimate deviated from the truth. Thus, although participants had the opportunity to accrue a large bonus, there was also a large penalty for error.

Procedure. Participants made nine estimates related to U.S. demography and commerce. The study employed a 2 (*Task type*: Bounded vs. Unbounded estimates) x 2 (*Estimation process*: Independent first vs. Discussion first) design. In the *Bounded* conditions participants made estimates on a limited scale by estimating the percentage of instances for which a particular event occurred (e.g. "What percentage of Philadelphians has no health insurance?"). In the *Unbounded* conditions participants made estimates on an unlimited scale by estimating the frequency with which a particular event occurred (e.g. "How many Philadelphians have no health insurance?"). This variation enabled us to manipulate participants' tendency to make extremely large estimation errors while keeping question topic constant (Table 2).

This task type variable was crossed with estimation process, which we manipulated by altering whether participants did or did not make independent estimates prior to engaging in discussion. The *Independent First* conditions followed the same procedure as Study 1, whereby participants made estimates for all items individually (Round 1), and then engaged in a discussion with a partner resulting in a new set of joint estimates (Round 2). In the *Discussion First* conditions participants made their estimates by discussing and reaching consensus on each estimate with a partner without first committing to independent estimates (Round 1). These participants also then had the opportunity to revise their estimates through discussion and offer a revised set of joint estimates (Round 2).

Variables. As in Study 1, we calculated the absolute difference of each estimate from the correct answer. For each item we used the mean and standard deviation of the absolute differences from the Round 1 *Independent First* conditions to standardize the absolute differences of dyad-level estimates (e.g., after discussion, averaging, etc.) so that performance could be compared across items. As in Study 1, we coded improvement in accuracy (reduced error) as positive. Estimates that were an order of magnitude away from (i.e. ten times greater or ten times smaller than) the correct answers were again classified as "extreme errors."

A coder blind to hypotheses viewed the videos of the discussion sessions for all conditions and recorded the estimates made by participants in the order in which they were verbalized. Additionally, for each item, participants reported their confidence that their individual or their jointly-produced estimate

fell within 10 percentage points of the answer on a five-point scale from *"Not at all confident"* to *"Extremely confident."*

Analytical approach. In our analyses we entered estimation error as an item-level (Level 1) variable and the two independent factor variables, *Estimation Process* (Independent: -1; Discussion: +1) and *Item type* (Bounded: -1; Unbounded: +1) as dyad-level (Level 2) variables. Our sample size was determined by the availability of participants during the time allocated to this study by our behavioral lab. Seven pairs of estimates were excluded from Round 1 analyses when dyad partners in the *Discussion First* condition (who were supposed to reach agreement on each estimate) did not enter the same estimate. Eleven additional pairs of estimates were excluded from Round 2 analyses for the same reason.

Results

The role of extreme errors. As in Study 1, participants in all conditions reduced their estimation errors from Round 1 to Round 2 (see Figure 2 for estimation errors on the two rounds of the task relative to the average error of an individual working alone). Most importantly, and replicating our Study 1 results, participants in the *Unbounded/Independent First* condition outperformed the level of accuracy available through simple averaging (b = .12, z = 3.68, p < 0.001, [0.06, 0.19]). By contrast, participants in the *Bounded/Independent First* condition produced discussed estimates that were no more accurate than the average of the earlier independent estimates (b = -.03, z = 0.96, p = .34). This difference in accuracy improvement relative to averaging differed between conditions (b = .08, z = 3.37, p = 0.001, [0.03, 0.12]).

When we examine the role of extreme errors in producing this between-condition difference we see that our manipulation did indeed result in many more large errors in the *Unbounded/Independent First* condition (223 or 36.4% of all estimates) than in the *Bounded/Independent First* condition (4, or 1% of all estimates). When we added a binary variable indicating whether at least one of the dyad members committed an extreme error in their initial estimate to the above analyses comparing the benefits of discussion between conditions, we see that the presence of an extreme error predicts accuracy improvement (b = .23, z = 2.29, p = 0.02, [0.03, 0.42]) and the effect of condition drops to non-significance (b = .02, z = 0.61, p = .54, [-.03, .07]). When we conduct a Monte Carlo Mediation analysis

(Selig & Preacher, 2008), we see that the 95% confidence interval for the indirect effect did not include zero (b = 0.06, [0.008, 0.114]). Thus, as hypothesized, the presence of extreme errors mediated the difference in accuracy improvement due to discussion between participants in the *Unbounded/Independent First* versus *Bounded/Independent First* conditions.

The role of independent estimates. The second question that we sought to address in Study 2 dealt with whether discussion should take place subsequent to or in absence of individual estimates. After all, even the Round 2 estimates in the *Unbounded/Independent First* condition contained a substantial number of extreme errors (154 or 25.3% of all estimates). It is possible that discussion prior to the first round of estimates would mitigate this problem and result in estimates that are substantially more accurate than either averaging or discussion following independent assessment.

Because participants in the *Discussion First* conditions made joint estimates during both rounds of the task, in order to compare their performance to the average of two independent estimates, we must use the average of the initial estimates produced in the *Independent First* conditions as the benchmark. To do this, for each item and separately in the *Bounded* and *Unbounded* conditions, we calculated the mean error that resulted from averaging two independent estimates and subtracted that constant from the errors that dyads produced through discussion in the two rounds of the *Discussion First* conditions (standardizing all errors as described above).

When we examine these differences on the first round of the task, we observe that when initial estimates were made through discussion for *Unbounded* items, those estimates were indistinguishable in accuracy from the average of two independent estimates, (b = .03, z = 0.65, p = .52, [-.07, .14]) and indeed were only slightly more accurate (b = 0.07, z = 1.92, p = .05, [-.001, .14]) than those of a single participant working alone. On the *Bounded* items the estimates reached in the *Discussion First* condition were significantly less accurate than those that would have resulted from averaging two independent estimates (b = -.24, z = 5.78, p < 0.001, [-.32, -.16]) and no different than those reached by a single individual (b = -.01, z = .20, p = .85, [-.06, .05]).

On the second round of the task, when participants in the *Discussion First* conditions had a second opportunity to discuss their estimates, they slightly outperformed averaging on the *Unbounded* items, although this difference was marginally significant (b = .08, z = 1.82, p = .07, [-.01, .18]). Second round estimates on the *Bounded* items remained substantially less accurate than the averaging benchmark (b = -.16, z = -3.79, p < .001, [-.24, -.08]). Thus, on *Bounded* items even after two rounds of discussion, participants who did not first make independent estimates were not able to match the accuracy of averaging two independent judgments.

Analysis of audio recordings. To better understand these results we examined the audio-recordings of participants' discussions in order to compare the range of first round estimates discussed by dyad members in the *Discussion First* condition, to the range of the independent first round estimates made by dyad members in the *Independent First* condition. This analysis revealed that one consequence of discussion is that it dramatically narrowed the range of considered estimates relative to the range of independent estimates for any given item. Specifically, in the *Independent First* condition the estimates of any two dyad members differed on average by 46.8% of the correct answer in the *Bounded* condition, and by 292% of the correct answer in the *Unbounded* condition. In contrast, in the *Discussion First* condition the range of estimates stated out loud was 26.3% of the correct answer for the *Bounded* items, and 41.1% of the correct answer for the *Unbounded* items. In other words, dyad members who discussed their estimates considered a range that was considerably narrower than did individuals working independently (*Bounded items*: b = .20, z = 6.10, p < 0.001, [.14, .27]; *Unbounded items*: b = 2.51, z = 4.67, p < .001, [1.5, 3.6]).

Considering a narrower range of estimates decreases the probability that the correct answer lies within that range. Specifically, in the *Independent First* condition the correct answer fell into the range created by the dyad members' two estimates 28.0% of the time for *Bounded* items, and 36.6% of the time for *Unbounded* items. By contrast, in the *Discussion First* condition the correct answer fell into the range of estimates that the dyad members considered only 13.1% and 6.1% of the time, for *Bounded* and *Unbounded* items respectively. Indeed, this low bracketing rate is similar to the bracketing rates observed

when a single individual is asked to offer multiple estimates (Herzog & Hertwig, 2009; Herzog & Hertwig, 2014). Thus, discussion in the absence of independent estimates severely reduced the likelihood that the range of estimates considered by participants included ("bracketed") the correct answer, curtailing the benefits of collaboration.

Discussion and prevention of extreme errors. Although dyads in the *Discussion First* condition considered a narrower range of estimates than two individuals working independently, we hypothesized that one benefit of placing discussion at the start of the judgment process was that it would prevent participants from making extreme errors that would hamper the accuracy benefits of collaboration. Indeed, when making estimates on an unbounded scale, participants making estimates independently made 223 out of 612 (36.4%) order of magnitude errors, which is higher than the 115 out of 396 (27.8%), that we observed in the *Discussion First* condition, $b = .39$, $z = 2.52$, $p < 0.01$, [0.09, 0.68]. (A similar pattern emerged on *Bounded* items, although of course the number of extreme errors was much lower (*Independent first:* 4 extreme errors vs. *Discussion first:* 2 extreme errors.)

Thus our data suggest that there is a trade-off between making independent estimates in order to increase the range of the judgments that are considered during discussion and placing discussion at the beginning of the judgment process to reduce the frequency of extreme errors. The benefit of early discussion in reducing extreme errors (which was only relevant for the *Unbounded* items) was offset on both *Bounded* and *Unbounded* items by reducing bracketing. The net effect of early discussion for *Bounded* items was detrimental because the reduction in bracketing was harmful and there was little available benefit from reducing extreme errors. The net effect of early discussion for *Unbounded* items was neither positive nor negative because the harm of reduced bracketing was offset by the benefits of reducing extreme errors.

General Discussion

Popular belief and even classical philosophy (Aristotle, 1995) hold that discussion enhances the quality of reasoning. Yet, most prior empirical research, with only a handful of exceptions, has shown that

discussed judgments are on par with the simple average of the relevant individual judgments. We show that in the case of dyadic judgment the effects of discussion are more complex. In our studies discussion aided accuracy when it enabled participants to prevent or eliminate egregious errors. When participants working on unbounded estimates engaged in discussion after offering independent estimates, discussion outperformed averaging. This effect emerged robustly across four participant samples and four sets of estimation items.

However, Study 2 also demonstrates a cost of discussion. Discussion in absence of independent judgments severely limited the range of considered estimates. Although this restriction of range took place for both types items used in our studies, it only proved to be highly detrimental to accuracy when the range of estimates was naturally bounded. When there were no large errors to reduce, discussed estimates severely underperformed averaging, even after two rounds of consideration. On unbounded items, although early discussion limited the commission of extreme errors, range restriction still prevented discussed estimates from outperforming averaging (Minson & Mueller, 2011).

Such restriction of range is a likely outcome of "anchoring and insufficient adjustment" (Tversky & Kahneman, 1974) of estimates relative to the first estimate spoken (Koehler & Beauregard, 2006). However, the magnitude of this effect is surprising. On bounded items the restriction of range made the discussed estimates of two people indistinguishable in accuracy from the estimate of a single individual. This suggests that even under incentivized conditions participants are not sufficiently attached to the estimates in their mind to resist the influence of peers.

To what extent are lay people aware of the costs and benefits of discussion highlighted by our research? To address this question we presented 208 respondents with two collaborative judgment scenarios: one regarding estimating the cost of an upcoming home remodel with their spouse (n =105), and another regarding forecasting the likely revenues generated by a new employee with a business partner (n = 103), both "unbounded" estimates. We then instructed participants to rank order seven judgment strategies in order of the accuracy of estimates that those strategies were likely to produce. Participants strongly endorsed strategies featuring discussion, preferring those to simple averaging of

individual inputs, or to various attempts to pick among the available estimates (see Appendix 1 for materials and full results). Furthermore, a majority of participants (58%) believed that having group members engage in two rounds of discussion would lead to better estimates than having them form independent judgments before discussion (vs. 50%, $p = .02$, two-tailed binomial test).

These intuitions are interesting because they suggest that participants have an appreciation for the value of discussion, but they do not recognize the costs associated with social influence. When we repeated the above survey using decision scenarios using a bounded scale, we obtained virtually identical results. Participants who considered estimating the percentage reduction in their energy bill from home improvements (n = 100) as well as participants who considered estimating the percentage revenue increase from a new business computer system (n = 108) continued to put their faith in discussion and again rated discussing the estimate twice more positively than making independent estimates prior to discussion (59%). Thus, individuals do not seem to recognize that unstructured discussion is more costly in some contexts than others.

In summary, our data suggest that determining whether to use discussion as part of a decision-making process engenders a series of tradeoffs. On one hand, discussion is by definition more time consuming than simple averaging. On the other, discussion is an effective tool for reducing very large, potentially catastrophic errors. However, discussion undertaken without first requiring decision-makers to voice independent judgments can also carry an accuracy cost. Managers, consumers and policy-makes would be wise to consider these trade-offs in planning decision processes.

How Does Discussion Benefit Accuracy?

An important question raised by our results is what accounts for the effectiveness of discussion in eliminating extreme errors? One possibility is that dyad members are able to recognize which member's preliminary answer looks more accurate and shift their focus to that estimate. This would be a pure "weighting" change. If dyad members start out with independent estimates (one or both of which might

be highly erroneous) and then adjust by giving greater weight to the more accurate of the two, there remains the question of how exactly such weighting is accomplished.

A simple explanation might involve the communication of greater confidence by the more accurate judge. However, when we regressed the standardized accuracy of independent estimates on participants' self-reported level of confidence in those estimates we saw no relationship between self-reported confidence and accuracy in either of the two studies (Study 1: b = .04, z = 0.84, p = .41, [-.05, .08]; Study 2: b = -.02, z = -0.49, p = .63, [-.10, .06]). The difference between own and partner's confidence was similarly unpredictive of accuracy. If the estimates of more confident participants were not more accurate, than communicating confidence would not have resulted in error reduction.

A second explanation may involve the fact that face-to-face discussion can give rise to greater accountability relative to working alone. Those who work alone do not have to justify their decisions to others. In contrast, dyad members face unique pressures toward accountability since there is no way to abstain from discussion. People may cope with the demands of accountability by ensuring their estimates have defensible reasoning behind them and so are more likely to be acceptable to others (Lerner & Tetlock, 1999). Indeed, research shows that making participants in a group accountable by making their responses identifiable increases participants' cognitive effort when engaging in a judgment task (Weldon & Gargano, 1988). Dyad members may put forth more cognitive effort in the course of thinking through initial estimates than members of crowds, or people working individually when their responses are confidential (Weldon & Gargano, 1988). Similarly, accountability may lead dyad members who are evaluating their own and their partner's estimate to put more cognitive effort into that evaluation. Future research should measure this mechanism directly, and further examine whether estimates in larger groups or crowds can also be improved with increased accountability.

A different set of explanations is based on the idea that rather than figuring out how to appropriately weight two highly erroneous estimates, discussion enables participants to more effectively represent the problem, which promotes use of valid cues and proper cue weights. For example, dyads might think about familiar cities comparable to Philadelphia in size when making estimates about

Philadelphians, or consider other large universities to estimate facts about the University of Michigan. Similarly, it may be the case that discussing estimates leads participants to appropriately scale the magnitude of their estimate by essentially adding or removing zeros. For example, in estimating the female population of Alaska, this may involve realizing that although Alaska has relatively fewer women than men, it is still a very large state.

Our data provide some evidence for this mechanism in that dyads in the *Discussion First* condition of Study 2 made fewer extreme errors on the first round of the task than dyads in the *Independent First* condition. However, these results are based on discussion that took place in absence of independent estimates, not discussion following independent estimates. In order for this "problem framing" mechanism to be the driver of error reduction through discussion after independent estimation, dyad members would have to abandon the individual, erroneous problem representation and develop a new, jointly considered representation.

Our data do not allow us to distinguish with precision which of the above mechanisms (or their combination), are responsible for the reduction of error through discussion. However, understanding this question is important in order to effectively design decision processes. Future research can address this question by developing a coding scheme to identify the frequency of various discussion processes based on a larger set of recorded data and testing their effect on error reduction by manipulating the process that participants follow.

Implications for Decision Processes

Our findings regarding the benefits of discussion should be considered in the context of previously tested techniques for combining the opinions of multiple judges. Most notably, the well-studied Delphi technique requires individuals to revise their judgments over multiple rounds based on the anonymous input of other group members and summary statistics regarding group consensus. Members of Delphi groups whose judgments appear to be extreme relative to the group average are often asked to

provide additional explanation in order to unearth faulty assumptions or more broadly share uniquely-held information (Rowe & Wright, 1999).

Dyadic discussion is different from the Delphi process in several respects. First, the presence of only two individuals makes anonymity impossible. Second, a Delphi group usually requires a moderator to solicit and combine the judgments provided by the members. This process is of necessity time consuming and logistically complex. Future research should address the relative costs (in terms of time and effort) versus benefits (in terms of greater accuracy) of the Delphi technique versus small group discussion, paying particular attention to the error distribution of the judgments in question.

Relatedly, the Nominal Group Technique (NGT) (Delbecq & Van de Ven, 1971; Graefe & Armstrong, 2011; Van de Ven & Delbecq, 1974) features a face-to-face discussion phase, which is both preceded and followed by individual estimates. The key difference between the NGT and the research described here is that the final product of the NGT is the average of the individual estimates that group members make after discussion, as opposed to a consensus estimate reached through discussion. We do not have a strong prediction about whether the NGT would perform better or worse than the process described here. However, it seems likely that similarly to our results, the NGT technique would prove particularly beneficial relative to simple averaging when egregious errors are likely.

While the Delphi method and the Nominal Group Technique both feature face-to-face discussion, other research has tested methods whereby individual judgments are combined according to some idiosyncratic weighting scheme without discussion. One example of this is the Judge Advisor System (JAS) (see Bonaccio & Dalal, 2006 for a review) wherein individuals decide how to weigh their own and a peer's estimate without discussion. A common finding in this literature is that the resulting judgments underperform the accuracy of the arithmetic mean because participants give too much weight to their own initial inputs, irrespective of accuracy. However, although the final judgments in JAS are not as good as averaging, they are still substantially more accurate than the judgments of a typical individual working alone. In other words, simple exposure to the judgments of others reduces error. Similar results have been obtained with larger groups as well (Farrell, 2011; Lorenz et al., 2011).

Finally, it is worth considering other methods of combining judgments that do not allow for any discretion by the decision makers in question. For example, although in our studies discussion outperformed simple averaging, there are other methods that are not as severely impacted by extreme errors. One possibility is to take the geometric mean instead of the arithmetic mean of the estimates in question because geometric means are known to severely curtail the influence of large over-estimates. Although in our particular data geometric means would have performed as well as if not better than discussion, this strategy relies on making risky assumptions about the shape of the error distribution. This mechanism would be detrimental when there are large under-estimates. The skew of the distribution might be a clue as to which errors are most likely to be outliers; but, in dyads, there is no opportunity to observe the skew.

A related approach is to take a median estimate instead of the mean. Although this option is not available to dyads, in larger groups the median can be a very effective way of reducing error. When we simulated groups of different sizes using our Study 1 data, we found that although discussion outperformed the mean of three estimates on 22 out of 27 items used in Study 1, it outperformed the median of three estimates on only 7 of the 27 items. Across all the items in our studies, the median of three independent estimates reduced absolute error by approximately 50% (compared to the average absolute error of individual estimates); discussion in dyads reduced absolute error by approximately 30%. Thus in cases when more than two judges can be solicited for independent estimates, taking the median of those judgments outperforms discussion by two individuals. Dyadic discussion following independent judgments however achieves a substantial portion of improvement available from the larger sample. Decision-makers must weigh whether the gain in accuracy using a larger sample justifies the increased investment of human resources.

Future Research

To the extent that discussion enables better accuracy by dyads, one might also consider how larger groups would perform as a result of discussion versus averaging. Our simulation demonstrated that

in domains rife with extreme errors even estimates based on averaging the individual judgments of a large number of participants do not typically outperform discussion by just two people. Thus although discussion by three or four individuals might reduce error even further, averaging is no longer an interesting benchmark for comparison. An intriguing alternative hypothesis is that discussion by a larger group of participants might result in *worse* accuracy than discussion by a dyad. Prior research has demonstrated that as groups get larger individuals in those groups put forth less effort (Liden, Wayne, Jaworski, & Bennett, 2004), speak less (Dunbar, Duncan, & Nettle, 1995), and may be less likely to ask for help if they feel confused (Mueller, 2012). Future research should address the specific range of group sizes at which discussion remains more effective than averaging and how it performs relative to the accuracy of taking the median.

Although our investigation focused on simple quantitative estimates, a similar conceptual analysis can be applied to more complex problems. Future research should examine whether discussion improves accuracy when collaborators make multifaceted decisions regarding the best candidate to hire, or the most effective military strategy to pursue, or the most exciting new product to invest in. Given that the effects of judgment errors can be amplified in complex systems (Silver, 2012), it may be particularly important to eliminate extreme errors through discussion when making judgments that will form assumptions for subsequent calculations.

Conclusion

By clarifying some of the tradeoffs inherent in understanding collaborative judgment, our results provide guidance to organizations and consumers who are interested in increasing judgment accuracy. First, if time and resources are committed to collaboration, compelling decision-makers to form a set of initial independent judgments can guard against the destructive assimilation effects caused by premature discussion. Second, and in contrast to most prior findings, on tasks with a high likelihood of extreme errors discussion following independent estimates consistently outperforms averaging. Given that error distributions of judgment tasks cannot be readily observed a priori, decision-makers would be wise to

consistently use a strategy of independent estimation followed by discussion to maximize accuracy across contexts.

Individuals firmly believe that discussion is beneficial to judgment accuracy even in light of empirical research the majority of which has concluded the opposite. Our results suggest that the lay psychologist may be on to something and that the variation in research results is not simply "noise" but the predictable result of identifiable factors. Even brief discussion by two people can offer dramatic benefits in the seemingly common cases when large errors are possible. When used prematurely, however, discussion may cost investors, managers, and consumers time, effort, and money, while offering no accuracy returns. Furthermore, people's preference for discussion seems insensitive to the value of initial independence or the potential for large errors. Taken together, our results suggest that although dyads and small groups can be wise like larger crowds, such wisdom is most likely achieved through an appreciation of the process and contextual factors that make collaboration effective.

References

Ancona, D. G., & Caldwell, D. F. 1992. Demography and Design: Predictors of New Product Team Performance. *Organization Science*, 3(3): 321-341.

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. 2000. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, 6(2): 130.

Aristotle, J. B. 1995. Complete works of Aristotle. *Ed. J. Barnes, Princeton, NJ*.

Armstrong, J. S. 2006. How to make better forecasts and decisions: avoid face-to-face meetings. . *Foresight*, 5: 3-8.

Beersma, B., & De Dreu, C. K. W. 2005. Conflict's consequences: Effects of social motives on postnegotiation creative and convergent group functioning and performance. *Journal of Personality and Social Psychology*, 89(3): 358-374.

Bonaccio, S., & Dalal, R. S. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2), 127-151.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4): 559-583.

De Dreu, C. K. W. 2006. When Too Little or Too Much Hurts: Evidence for a Curvilinear Relationship Between Task Conflict and Innovation in Teams. *Journal of Management*, 32(1): 83-107.

De Dreu, C. K. W., Nijstad, B. A., & van Knippenberg, D. 2008. Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review*, 12(1): 22-49.

De Dreu, C. K. W., & West, M. A. 2001. Minority dissent and team innovation: The importance of participation in decision making. *Journal of Applied Psychology*, 86(6): 1191-1201.

Delbecq, A. L., & Van de Ven, A. H. 1971. A group process model for problem identification and program planning. *The Journal of Applied Behavioral Science*, 7(4): 466-492.

Deutsch, M., & Gerard, H. B. 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3): 629.

Dunbar, R. I. M., Duncan, N. D. C., & Nettle, D. 1995. Size and structure of freely forming conversational groups. *Human Nature*, 6(1): 67-78.

Efron, B., & Tibshirani, R. J. 1994. An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

Elsbach, K. D., & Kramer, R. M. 2003. Assessing creativity in hollywood pitch meetings: evidence for a dual-process model of creativity judgments. *Academy of Management Journal*, 46(3): 283-301.

Farrell, S. 2011. Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences*, 108(36), E625-E625.

Galton, F. 1907. Vox populi. *Nature*, 75: 450-451.

Gigone, D., & Hastie, R. 1997. Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, 121(1): 149.

Graefe, A., & Armstrong, J. S. 2011. Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting*, 27(1): 183-195.

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(3), 305-326.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. Psychological Science, 20(2), 231–237.

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. Trends in Cognitive Sciences, 18(10), 504–506.

Hogarth, R. 1978. A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21(1): 40-46.

Klugman, S. F. 1947. Group and individual judgments for anticipated events. *The Journal of Social Psychology*, 26(1): 21-28.

Knight, H. C. 1921. A comparison of the reliability of group and individual judgments, *Unpublished master's thesis*. Columbia University.

Koehler, D. J., & Beauregard, T. A. 2006. Illusion of confirmation from exposure to another's hypothesis. *Journal of Behavioral Decision Making*, 19(1): 61-78.

Kurtzberg, T. R. 2005. Feeling Creative, Being Creative: An Empirical Study of Diversity and Creativity in Teams. *Creativity Research Journal*, 17(1): 51-65.

Larrick, R. P., & Soll, J. B. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, 52(1): 111-127.

Lerner, J. S., & Tetlock, P. E. 1999. Accounting for the effects of accountability. *Psychological Bulletin*, 125(2): 255-275.

Liden, R. C., Wayne, S. J., Jaworski, R. A., & Bennett, N. 2004. Social Loafing: A Field Investigation. *Journal of Management*, 30(2): 285-304.

Lobo, M. S., & Yao, D. 2010. Human judgement is heavy tailed: Empirical evidence and implications for the aggregation of estimates and forecasts. *Fontainebleau: INSEAD*.

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

Lorge, I., Fox, D., Davitz, J., & Brenner, M. 1958. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological Bulletin*, 55(6): 337.

Makridakis, S., & Winkler, R. L. 1983. Averages of forecasts: Some empirical results. *Management Science*, 29(9): 987-996.

Minson, J. A., Liberman, V., & Ross, L. 2011. Two to Tango Effects of Collaboration and Disagreement on Dyadic Judgment. *Personality and Social Psychology Bulletin*, 37(10): 1325-1338.

Minson, J. A., & Mueller, J. S. 2012. The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science*, 23(3): 219-224.

Minson, J. A., & Mueller, J. S. 2013. Groups Weight Outside Information Less Than Individuals Do, Although They Shouldn't Response to Schultze, Mojzisch, and Schulz-Hardt (2013). *Psychological Science*, 24(7): 1373-1374.

Mooney, C. Z., Duval, R. D., & Duvall, R. 1993. *Bootstrapping: A nonparametric approach to statistical inference*: Sage.

Mueller, J. S. 2012. Why individuals in larger teams perform worse. *Organizational Behavior and Human Decision Processes*, 117(1): 111-124.

Reagan-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(2): 246-270.

Rowe, G., & Wright, G. 1999. The Delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, 15(4): 353-375.

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(1): 24-36.

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. 2013. Groups Weight Outside Information Less Than Individuals Do Because They Should Response to Minson and Mueller (2012). *Psychological Science*, 24(7): 1371-1372.

Silver, N. 2012. *The signal and the noise: Why so many predictions fail-but some don't*. NY, NY: Penguin.

Simmons, J., Nelson, L., & Simonsohn, U. 2012. A 21 word solution. *Available at SSRN 2160588*.

Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. 2011. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1): 1-15.

Sniezek, J. A., & Henry, R. A. 1989. Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, 43(1): 1-28.

Sunstein, C. R. 2006. *Infotopia: How many minds produce knowledge*: Oxford University Press.

Surowiecki, J. 2005. *The wisdom of crowds*: Random House Digital, Inc.

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

Van de Ven, A. H., & Delbecq, A. L. 1974. The effectiveness of nominal, Delphi, and interacting group decision making processes. *Academy of management Journal*, 17(4): 605-621.

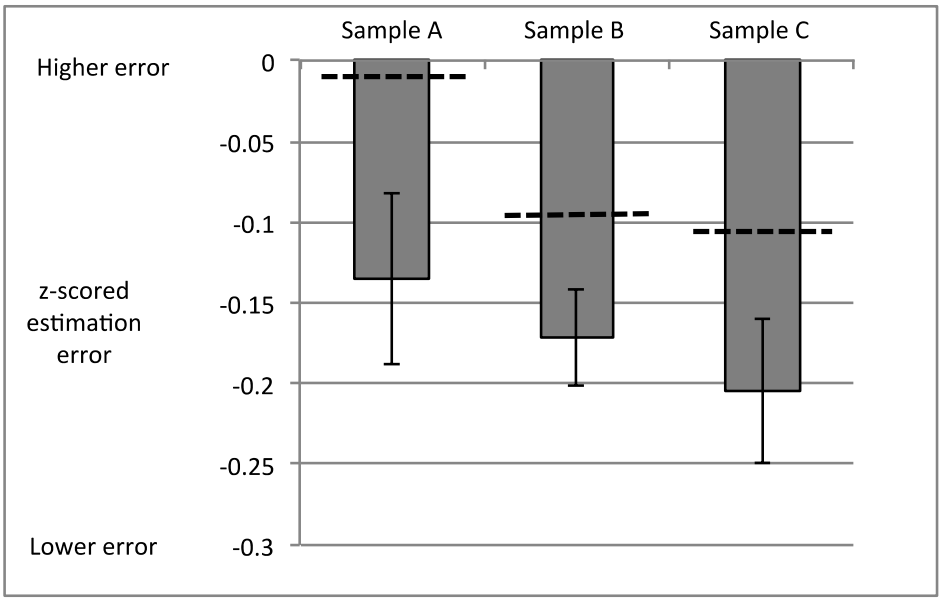
Vul, E., & Pashler, H. 2008. Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19(7): 645-647.

Weldon, E., & Gargano, G. M. 1988. Cognitive loafing: The effects of accountability and shared responsibility on cognitive effort. *Personality and Social Psychology Bulletin*, 14(1): 159-171.

Table 1: Estimation items used in Study 1.

Table 2: Estimation items used in Study 2.

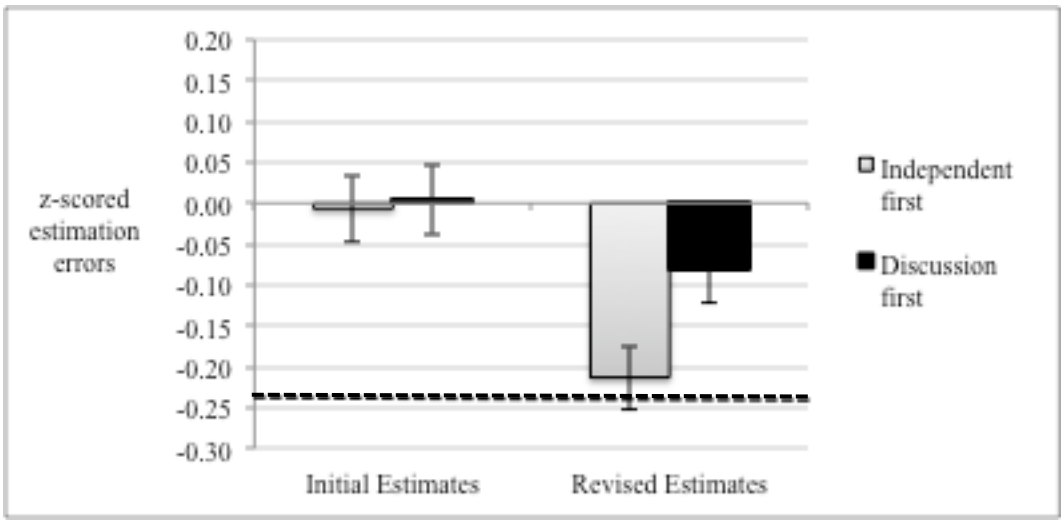
Figure 1: Estimation errors for each of the samples in Study 1. The length of each bar represents the reduction in absolute error achieved through discussion compared to the average individual's absolute error. The dashed lines represent the errors that would have been possible through averaging. Error reduction is scaled using the standard deviation of the individual absolute error.



Note: Standard error bars represent difference between accuracy of discussed judgments and accuracy available through averaging of dyad members' initial estimates.

Figure 2: Errors of initial and revised estimates. The length of each bar represents the reduction in absolute error compared to the average absolute error of initial estimates (either independent or from discussion). Dashed lines represent the error that would have resulted from averaging the independent Round 1 estimates of dyad members in the Independent First condition. Error reduction is scaled using the standard deviation of the absolute error of initial estimates.

a.) Bounded items



b.) Unbounded items