*Journal of Behavioral Decision Making*, *J. Behav. Dec. Making*, **28**: 250–261 (2015)
Published online 3 November 2014 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1843

# The Power of Social Influence on Estimation Accuracy

BURCU GÜRÇAY*, BARBARA A. MELLERS and JONATHAN BARON
*Department of Psychology, University of Pennsylvania, Philadelphia, PA USA*

ABSTRACT

Research shows that crowds can provide more accurate estimates of uncertain quantities than individuals (Surowiecki, 2004). But little is known about how to organize crowd members to maximize accuracy. When should crowd members work independently, and when should they work collaboratively? We examined the effects of social influence on estimation accuracy, consensus, and confidence. Participants first made independent estimates of uncertain quantities, such as the percentage of U.S. deaths due to heart attacks or the height of the tallest building. Then, in some conditions, they interacted with others online. After the discussion, they made second independent estimates. Social interaction improved accuracy. Despite well-known problems with groups, such as herding and free riding, discussion resulted in more accurate estimates and greater consensus relative to independent estimates. We offer a simple model that describes the process by which group discussion improves the estimates of uncertain quantities. Copyright © 2014 John Wiley & Sons, Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web-site.

KEY WORDS wisdom of crowds; group judgment; group versus individual incentives; group decision making

## INTRODUCTION

Groups have the potential to produce decisions that are far better than those based on individuals working alone (Fraidin, 2004; Hertel, Kerr, & Messé, 2000; Laughlin, Hatch, Silver, & Boh, 2006; Michaelson, Watson, & Black, 1989; Tindale & Larson, 1992). Process gains have been demonstrated when groups are cohesive, have strong productivity norms, and share a mental model of the task (Kerr & Tindale, 2004; Levine & Moreland, 1990). Groups can motivate individuals who wish to perform well in the presence of others (Hertel et al., 2000).

However, history and research show that groups also can make decisions that are far worse than those of individuals (Branson, Steele, & Sung, 2010) even in the presence of diversity, strong leadership, and unlimited time (Kerr & Tindale, 2004; Tindale, Smith, Thomas, Filkins, & Sheffey, 1996; Wittenbaum & Stasser, 1996). The suboptimality of group decisions has been attributed to several factors, including groupthink (Janis, 1982), social loafing (Latané, Williams, & Harkins, 1979), conformity (Asch, 1951; Sherif, 1936), misplaced competition (McGrath, 1984), and excessive focus on common information (Gigone & Hastie, 1993; Larson, Foster-Fishman, & Keys, 1994; Stasser, Taylor, & Hanna, 1989; Stasser & Titus, 1987; Wittenbaum & Park, 2001).

### The Delphi method for group estimates of uncertain quantities

To overcome such failings, Dalkey and Helmer (1963) proposed a structured discussion technique known as the Delphi method (Brown, 1968; Sackman, 1974). Group members begin with independent judgments. These judgments and rationales are then given to all group members by a moderator who ensures anonymity. After considering others' judgments, members can revise their opinions, and the process is iterated until consensus is reached. Rowe and Wright (1999) examined the effectiveness of the Delphi method and found that it outperformed independent groups in 23 studies with two ties and standard interacting groups in five studies with two ties. The success of the Delphi procedure has been demonstrated across many domains. The method has yielded better decisions than uncontrolled group discussions in predictions of social technological events (Kaplan, Skogstad, & Girshick, 1949), controversial and emotionally charged problems (Van de Ven & Delbecq, 1974), and estimates of minimum expenditures of tourists in Catalonia, Spain (Landeta, 2006).

What makes the Delphi method work better than others? The Delphi method requires four components: anonymity, iteration, statistical aggregation of the group response, and controlled feedback (Rowe & Wright, 1999). Anonymity is achieved through the use of private questionnaires and may reduce social pressures. Iteration allows people to change their minds, and statistical aggregation is usually conducted using a mean or median. Controlled feedback refers to the method of aggregating opinions from the previous round and giving it to participants in the next round. Feedback could be a set of individual estimates, summary statistics, or a set of arguments. Rowe, Wright, and McColl (2005) found that feedback in the form of statistics or arguments improved accuracy over no feedback.

Given these findings, we were surprised by a recent study that claimed estimates from social groups that received statistical feedback were no more accurate than estimates based on independent groups without feedback (Lorenz, Rauhut, Schweitzer, & Helbing, 2011). Lorenz et al. asked participants to make estimates of uncertain quantities, such as the number of murders in Switzerland in 2006. The study had three conditions. In one condition, participants were given a summary statistic of their group members' estimates, and in the other condition, they received the full distribution of individual estimates over five rounds. In a third condition, participants made estimates over five rounds with no new information. All participants answered a third of

*Correspondence to: Burcu Gürçay, Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: bgurcay@sas.upenn.edu

the questions in the summary statistic condition, another third in the full information condition, and another third in the no new information condition. Lorenz et al. argued that the accuracy of the interacting groups with feedback was no better than that of the aggregation of individuals without feedback who worked alone.

We devised a simple estimation task inspired by Lorenz et al. (2011) to understand the following: (i) the effects of social interaction; (ii) the effects of incentives; and (iii) the process by which social interaction helps or hurts. Why were the groups whose members interacted no better than those whose members worked independently? Perhaps statistical feedback was simply not enough. Greater accuracy may have required rationale feedback in addition to statistical feedback. Best (1974) found greater accuracy in a Delphi group that received both statistical and rationale feedback, compared with a group that received just statistical feedback.

Another possibility is that groups were not incentivized to work together. Lorenz et al. did not manipulate incentives, but their results might have differed if participants had been incentivized to maximize group accuracy rather than individual accuracy (Bonner, Hastie, Sprinkle, & Young, 2000; DeMatteo, Eby, & Sundstrom, 1998; Farr, 1976). If participants in the Lorenz et al. study were insufficiently motivated, they might not have bothered to revise their estimates at all.

Two prominent theories have been offered to explain the effects of incentives. One emphasizes equity and focuses on differences among group members (Adams, 1963, 1965). Individual talents and efforts are recognized and rewarded. The other emphasizes equality and focuses on similarities among group members (Deutsch, 1949). In general, equity incentives promote competition, and equality incentives promote cooperation. Individual incentives often improve group speed because of competition, whereas group incentives improve group accuracy because of collaboration (Beersma et al., 2003). Some argue that the combination of individual and group incentives is the best of the both worlds (De Dreu, Nijstad, & Van Knippenberg, 2008; DeMatteo et al., 1998; Heneman & von Hippel, 1995; Kozlowski & Ilgen, 2006; Welbourne & Gomez Mejia, 1995).

### Understanding the process

In the study that follows, we asked participants to provide confidence ratings along with each of their estimates and reasoned that these ratings might be used to infer individual expertise. Some studies, however, suggest that the connection between confidence and accuracy is weak (Phillips, 1999; Rowe & Wright, 1996). Nonetheless, there is evidence to suggest that less confident participants are more likely than more confident participants to seek advice and change their opinions (Cooper, 1991).

We propose that subject $i$'s second-round estimate of an uncertain quantity on question $j$ ($e_{2ij}$) is a weighted average of his or her first-round estimate ($e_{1ij}$) and the median of the group's first-round estimates. Person $i$'s estimate is weighted by his or her confidence, expressed as the number of people who that person believes are less accurate ($c_{1ij}$). The model is

$$\hat{e}_{2ij} = \frac{C_{1ij}}{n} \times e_{1ij} + \frac{(n - c_{1ij})}{n} \times median\left(e_{11j}, \ldots, e_{1nj}\right)$$

where $n$ is the number of people in the group. This model

Table 1. Estimation questions, answers, and median accuracy scores

| Questions | Correct answers | Median accuracy scores | | | | | |
|---|---|---|---|---|---|---|---|
| | | II | | GI | | GG | |
| | | First round | Second round | First round | Second round | First round | Second round |
| 1 What percentage of the U.S. land does Texas occupy? | 7.08 | 6.9 | 7 | 4.9 | 2.9 | 2.9 | 1.9 |
| 2 What percentage of Americans was not medically insured in 2010? | 16.4 | 13.6 | 13.6 | 13.6 | 8.6 | 13.6 | 13.6 |
| 3 What percentage of the deaths in the U.S. was caused by heart disease (e.g., heart attacks and strokes) in 2008? | 24.9 | 10 | 10 | 9.9 | 8 | 11.1 | 13.1 |
| 4 What percentage of the people in the U.S. is Catholic? | 25 | 7.5 | 5 | 10 | 7 | 10 | 5 |
| 5 What is the divorce rate in America for first marriages? | 41 | 9 | 9 | 9 | 9 | 9 | 9 |
| 6 What percentage of Americans did not pay income taxes in 2009 because they did not have the required minimum income? | 49.5 | 29.5 | 29.5 | 33 | 33 | 30 | 32 |
| 7 What percentage of Americans has a pet? | 63 | 13 | 16 | 23 | 25 | 23 | 23 |
| 8 What was the inflation rate in the U.S. in March 2012? | 2.7 | 1.3 | 1.5 | 2.3 | 0.8 | 2.3 | 0.8 |
| 9 What is the length of the world's longest music video in minutes by 2010? | 39.52 | 28 | 26.8 | 25 | 15 | 25 | 15 |
| 10 How tall is the tallest dog ever in inches? | 43 | 17 | 17 | 17 | 17 | 13 | 17 |
| 11 What's the maximum testing speed of the fastest train in the world in mph? | 236 | 75 | 86 | 86 | 64 | 86 | 64 |
| 12 How tall is the highest man-made building in feet? | 2723 | 1770 | 1673 | 1723 | 1148 | 2093 | 948 |
| 13 What's the equatorial circumference of the Earth in miles? | 24 901 | 24 977 | 24 975 | 20 500 | 7099 | 20 151 | 9901 |
| 14 What was the U.S. gross domestic product per capita (income per person) in 2010 according to World Bank? (in international dollars) | 47 184 | 17 184 | 12 816 | 17 000 | 12 184 | 17 184 | 17 184 |
| 15 What is the total area of the U.S. in square miles? | 3 718 691 | 3 624 191 | 3 668 691 | 3 678 691 | 3 672 441 | 3 698 691 | 3 679 191 |
| 16 What was the population of London in July 2010? | 7 825 200 | 4 825 200 | 4 825 200 | 4 825 200 | 2 225 200 | 4 825 200 | 2 174 800 |



DOI: 10.1002/bdm

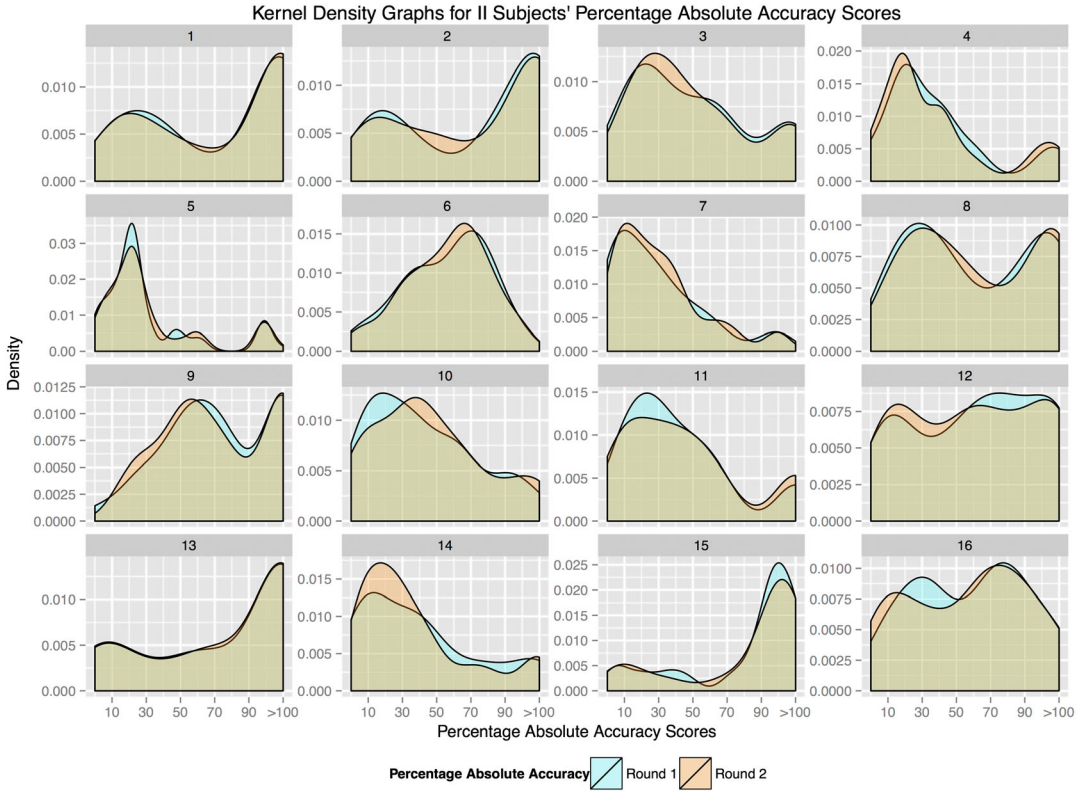


Figure 1. Kernel density graphs for participants' percentage absolute accuracy scores in the II condition. Zero is the perfect score, so less accurate estimates are values further to the right in each graph. Blue areas show first-round estimates, orange shows second-round estimates, and green shows the overlap. There is no systemic difference between first-round and second-round estimates

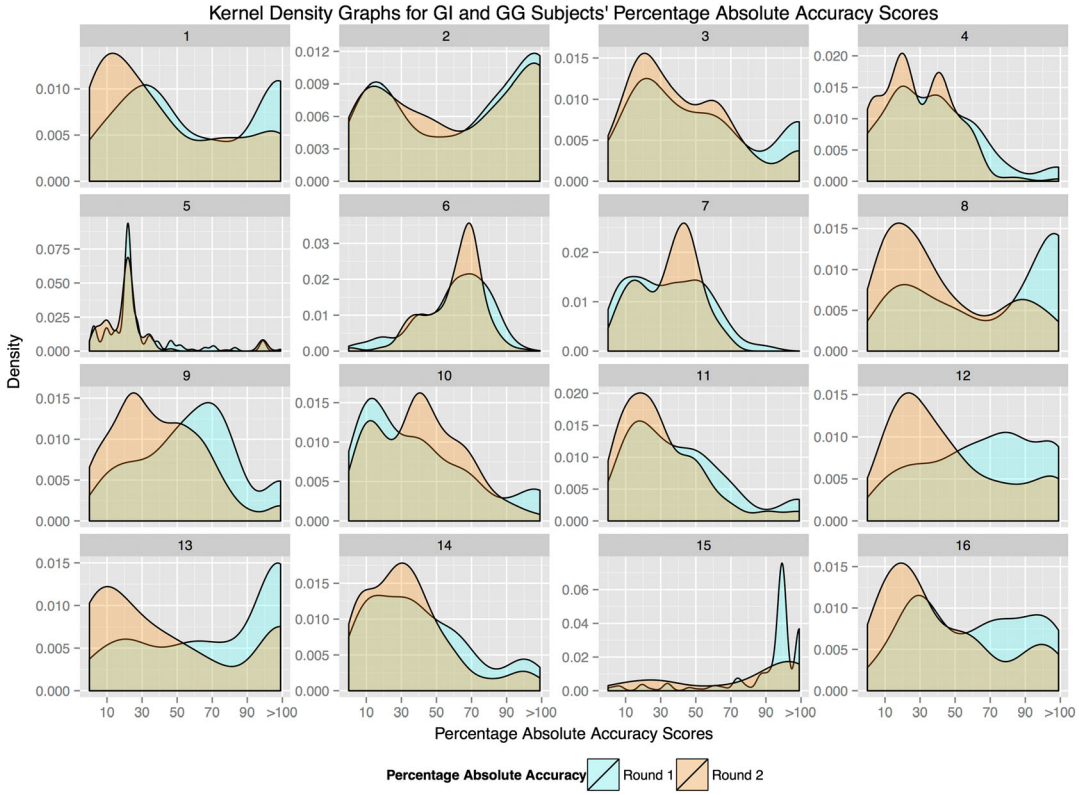


Figure 2. Kernel density graphs for participants' percentage absolute accuracy scores in the GI and GG conditions pooled together for each question and round. Zero is the perfect score, so less accurate estimates are values further to the right in each graph. Blue areas show first-round estimates, orange shows second-round estimates, and green shows the overlap. First-round estimates tend to be less accurate (more blue regions on the right), and second-round estimates tend to be more accurate (more orange regions on the left)



**DOI**: 10.1002/bdm

predicts that group discussion (represented as *median* $(e_{11j}, \ldots, e_{1nj})$) can increase both accuracy and consensus.

We hypothesize that individuals who are given the opportunity to discuss the question will make more accurate estimates and reach greater consensus than individuals working alone. In addition, group incentives should improve accuracy relative to individual incentives.

## METHOD

The purpose of our study was to understand crowd wisdom and reconcile the differences between Lorenz's results and previous results. We focused on three questions. Does social influence improve accuracy? Do group incentives improve accuracy? And how can we model the social influence process? Undergraduates from the University of Pennsylvania served as participants. They were given 16 questions shown in Table 1 that were selected from a pilot study of 54 questions. Questions were selected to be neither too hard nor too easy. Topics ranged from demographics to world records. Seven questions were percentages, and nine were open-ended and non-negative.

There were three conditions. In the first condition, participants worked independently with individual payment incentives (II). Participants made two rounds of estimates for each question. Payments were based on individual accuracy (i.e., the absolute difference between the estimate and truth) of their second-round estimates. In the second and third conditions, participants made initial estimates independently. Then they discussed the question, their first-round estimates, and their reasons for making these estimates online in a group of 8–10 people to increase their accuracy, although all group members were in the same room. Each member submitted a second independent estimate after a 90-second discussion for each question. The second condition (GI) paid participants according to individual accuracy, defined as the absolute difference between the second estimate and truth. The third condition (GG) paid participants according to group accuracy, defined as the absolute difference between the median of the individual estimates and truth.[1] There were from 92 to 96 participants in each condition.

A comparison of the second round of II responses with the second round of GI responses provides a between-subject test of whether social influence improves accuracy, increases confidence, and increases consensus when group members were individually incentivized (i.e., paid for their own accuracy). A comparison of the GI and GG conditions provides a between-subject test of whether group incentives produce more accurate estimates than individual incentives.

In all conditions, participants sat in a cubicle in front of a computer and were unable to see what others were doing. They were not allowed to communicate or use the Internet (or any mobile device) to gather information. In the II

[1]An alternative accuracy measure for continuous questions is the absolute value of the log of the ratio of the estimate to the correct answer. We analyzed the continuous questions by using this alternative dependent variable, and we observed no differences in results.

Table 2. Median accuracy scores and Wilcoxon signed-rank sum test statistics

| | II first | II second | *W* value (*N*) | *p* value | GG first | GG second | *W* value (*N*) | *p* value | GI first | GI second | *W* value (*N*) | *p* value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.9 | 7 | 362 (92) | 0.87 | 2.9 | 1.9 | 499 (70) | **<0.01** | 4.9 | 2.9 | 333 (65) | **<0.01** |
| 2 | 13.6 | 13.6 | 762 (92) | 0.82 | 13.6 | 13.6 | 767 (65) | **0.02** | 13.6 | 8.6 | 1143 (76) | **0.05** |
| 3 | 10 | 10 | 552 (92) | 0.45 | 11.1 | 13.1 | 851 (69) | **0.02** | 9.9 | 8 | 299 (55) | **<0.01** |
| 4 | 7.5 | 5 | 1076 (92) | 0.98 | 10 | 5 | 862 (74) | **<0.01** | 10 | 7 | 1074 (76) | **0.02** |
| 5 | 9 | 9 | 573 (92) | 0.20 | 9 | 9 | 376 (46) | **0.04** | 9 | 9 | 707 (64) | **0.01** |
| 6 | 29.5 | 29.5 | 629 (92) | 0.12 | 30 | 32 | 1461 (74) | 0.66 | 33 | 33 | 974 (65) | 0.26 |
| 7 | 13 | 16 | 706 (92) | 0.75 | 23 | 23 | 1410 (74) | 0.55 | 23 | 25 | 1276 (66) | 0.86 |
| 8 | 1.3 | 1.5 | 860 (92) | 0.77 | 2.3 | 0.8 | 445 (77) | **<0.01** | 2.3 | 0.8 | 469 (79) | **<0.01** |
| 9 | 28 | 26.8 | 882 (92) | 0.67 | 25 | 15 | 405 (76) | **<0.01** | 25 | 15 | 567 (75) | **<0.01** |
| 10 | 17 | 17 | 906 (92) | 0.92 | 13 | 17 | 1135 (68) | 0.41 | 17 | 17 | 1476 (74) | 0.68 |
| 11 | 75 | 86 | 711 (92) | 0.58 | 86 | 64 | 946 (81) | **<0.01** | 86 | 64 | 1011 (79) | **<0.01** |
| 12 | 1770 | 1673 | 549 (92) | 0.07 | 2093 | 948 | 629 (82) | **<0.01** | 1723 | 1148 | 653 (73) | **<0.01** |
| 13 | 24 977 | 24 975 | 812 (92) | 0.29 | 20 151 | 9901 | 1098 (84) | **<0.01** | 20 500 | 7099 | 388 (75) | **<0.01** |
| 14 | 17 184 | 12 816 | 648 (92) | 0.27 | 17 184 | 17 184 | 1145 (75) | 0.07 | 17000 | 12 184 | 936 (78) | **<0.01** |
| 15 | 3 624 191 | 3 668 691 | 732 (92) | 0.65 | 3 698 691 | 3 679 191 | 1092 (82) | **<0.01** | 3 678 691 | 3 672 441 | 1771 (83) | 0.48 |
| 16 | 4 825 200 | 4 825 200 | 896 (92) | 0.99 | 4 825 200 | 2 174 800 | 1061 (82) | **<0.01** | 4 825 200 | 2 225 200 | 564 (79) | **<0.01** |

Statistically significant questions are shown in bold font.





**DOI**: 10.1002/bdm

Table 3. Median accuracy scores and Mann–Whitney *U* test statistics

| | II | GG | *U* value ($N_1$, $N_2$) | *p* value | GI | *U* value ($N_1$, $N_2$) | *p* value |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 1.9 | 6090 (92, 91) | <**0.01** | 2.9 | 6230 (92, 92) | <**0.01** |
| 2 | 13.6 | 14 | 4894 (92, 91) | **0.03** | 9 | 5078 (92, 93) | **0.01** |
| 3 | 10 | 13 | 4148 (92, 92) | 0.59 | 8 | 5322 (92, 94) | <**0.01** |
| 4 | 5 | 5 | 4868 (92, 92) | **0.04** | 7 | 4470 (92, 93) | 0.34 |
| 5 | 9 | 9 | 4449 (92, 92) | 0.27 | 9 | 5080 (92, 94) | **0.01** |
| 6 | 29.5 | 32 | 4087 (92, 92) | 0.66 | 33 | 3655 (92, 94) | 0.97 |
| 7 | 16 | 23 | 3459 (92, 91) | 0.98 | 25 | 3134 (92, 93) | 0.99 |
| 8 | 1.5 | 0.8 | 5490 (92, 92) | <**0.01** | 0.8 | 5872 (92, 94) | <**0.01** |
| 9 | 26.8 | 15 | 6492 (92, 92) | <**0.01** | 15 | 6723 (92, 93) | <**0.01** |
| 10 | 17 | 17 | 4667 (92, 92) | 0.09 | 17 | 4271 (92, 93) | 0.46 |
| 11 | 86 | 64 | 5254 (92, 92) | <**0.01** | 64 | 5249 (92, 93) | <**0.01** |
| 12 | 1673 | 948 | 5169 (92, 92) | <**0.01** | 1148 | 5215 (92, 94) | <**0.01** |
| 13 | 24975 | 9901 | 5816 (92, 92) | <**0.01** | 7099 | 5888 (92, 93) | <**0.01** |
| 14 | 12 816 | 17 184 | 3801 (92, 92) | 0.88 | 12 184 | 4802 (92, 94) | 0.08 |
| 15 | 3 668 691 | 3 679 191 | 4885 (92, 92) | **0.04** | 3 672 441 | 4305 (92, 94) | 0.52 |
| 16 | 4 825 200 | 2 174 800 | 5131 (92, 92) | <**0.01** | 2 225 200 | 5359 (92, 94) | <**0.01** |

Statistically significant questions where GG and GI participants were more accurate appear in bold font.

condition, participants made two independent estimates for each question. They also rated their confidence on a scale from 0 to 9, where 0 = *not at all confident* and 9 = *extremely confident*. The order of the questions was randomized within and across conditions.

In the GI and GG conditions, participants made two rounds of estimates for each question. In the first round, estimates were independent. Then participants had 90 seconds to share information by writing their opinions on a chat page that showed a table of each group member's initial estimates and confidence ratings. The confidence question asked, "How many other people in your group (e.g., 0–9) do you think will do worse than you on this question?" In the second round, participants could revise their estimates and assign a second confidence rating.

All participants were paid $10 for their participation, and they could earn additional money for accuracy. Accuracy payments were based on the second-round individual estimates in the II and GI conditions and the second-round median estimates in the GG condition. If an estimate fell within ±10% of the correct answer, the reward was $1 and was paid to each individual (in the II condition) or each group member (in the GI condition). If an estimate fell between ±11% and ±20% of the correct answer, the reward was 50 cents. If an estimate fell between ±21% and ±40% of the correct answer, the reward was 25 cents. Larger absolute differences were not rewarded. There were 16 questions, so participants could earn $16, plus $10 for participation. Actual payments were an average of $3.71, $4.60, and $4.70 for accuracy plus $10 for participation in the II, GI, and GG conditions, respectively.

## RESULTS

### Social influence improves accuracy

We examined the distributions of estimates for each group on each question in each round of each condition. Distributions were skewed, so the analyses that follow are non-parametric. We used Mann–Whitney *U* tests for between-subject comparisons of means and Wilcoxon signed-rank sum tests for within-subject comparisons of means.

Figures 1 and 2 show the distributions of percentage absolute accuracy scores, defined as the absolute value of the difference between an estimate and truth divided by truth and then multiplied by 100. In Figure 2, these scores are averaged over questions and rounds in both the GI and GG conditions.[2] Zero is a perfect score. Less accurate estimates are further to the right. Blue areas represent first-round estimates, orange areas show second-round estimates, and green areas show the overlap. In Figure 2, first-round estimates are less accurate, and second-round estimates are more accurate and closer to zero, while Figure 1 does not show any difference between first-round and second-round estimates. Social interaction was beneficial to accuracy.

We had two ways of testing the effects of social influence on accuracy. The first was a within-subject comparison. Table 2 shows median accuracy scores for the first-round and second-round estimates of participants in the II, GG, and GI conditions. In the II condition, second-round estimates were no more accurate than first-round estimates in any of the questions. In both of the group conditions, second-round estimates were significantly more accurate than first-round estimates for 75% of questions.[3] Statistically significant questions are shown in bold font. Tests were one-tailed Wilcoxon signed-rank sums. Discussion improved accuracy within individuals from the first to second rounds.

The second way to test the effect of social influence was a between-subject comparison of first-round estimates from the II condition and second-round estimates from the GI or GG conditions. Even with different participants in treatment and control conditions, social influence improved accuracy. Table 3 shows median individual accuracy scores for the II, GG, and GI conditions, where scores in the GG and GI conditions are second-round estimates. Statistically significant

[2]The graphs report percentage accuracy scores, and scores that were at 100% and higher were grouped to display the data more clearly.

[3]The binomial probability of having 12 successes (assuming $p = 0.5$) is 0.0002.



**DOI**: 10.1002/bdm

Table 4. Estimate standard deviations and Pitman–Morgan test statistics for participants' estimates

| | II first | II second | *r* | *df* | *t* | *p* | GG first | GG second | *r* | *df* | *t* | *p* | GI first | GI second | *r* | *df* | *t* | *p* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 7 | 0.84 | 90 | 0.6 | 0.58 | 6 | **4** | 0.53 | 79 | 5.1 | <0.01 | 6 | **3** | 0.56 | 80 | 8.1 | <0.01 |
| 2 | 14 | 16 | 0.87 | 90 | 1.4 | 0.16 | 15 | **11** | 0.48 | 82 | 2.7 | <0.01 | 14 | 13 | 0.66 | 91 | 1.2 | 0.23 |
| 3 | 19 | **18** | 0.96 | 90 | 2.1 | 0.04 | 20 | **12** | 0.53 | 87 | 5.8 | <0.01 | 16 | **13** | 0.6 | 69 | 2.2 | 0.03 |
| 4 | 15 | 14 | 0.87 | 90 | 0.4 | 0.71 | 14 | **7** | 0.42 | 90 | 7 | <0.01 | 12 | **9** | 0.46 | 92 | 2.9 | <0.01 |
| 5 | 17 | 18 | 0.89 | 90 | 0.9 | 0.39 | 13 | **6** | 0.51 | 82 | 8.7 | <0.01 | 14 | 15 | 0.35 | 91 | 0.6 | 0.55 |
| 6 | 15 | **13** | 0.84 | 90 | 2.5 | 0.01 | 11 | **8** | 0.68 | 89 | 4.4 | <0.01 | 10 | **8** | 0.54 | 90 | 3.2 | <0.01 |
| 7 | 20 | 20 | 0.84 | 90 | 0.5 | 0.63 | 17 | **11** | 0.53 | 90 | 5.3 | <0.01 | 18 | **11** | 0.59 | 82 | 5.6 | <0.01 |
| 8 | 10 | 12 | 0.26 | 90 | 1.7 | 0.09 | 8 | **2** | 0.06 | 90 | 20.1 | <0.01 | 21 | **1** | 0 | 91 | 74 | <0.01 |
| 9 | 7498 | 8329 | 1.00 | 90 | 38.8 | <0.01 | 380 | **21** | 0.21 | 89 | 87.4 | <0.01 | 1088 | **19** | 0.05 | 82 | 264 | <0.01 |
| 10 | 17 | 19 | 0.87 | 90 | 2.1 | 0.04 | 144 | **73** | 0.68 | 90 | 9.4 | <0.01 | 3088 | **11** | −0.29 | 90 | 1356 | <0.01 |
| 11 | 730 | 743 | 0.95 | 90 | 0.5 | 0.61 | 186 | **79** | 0.17 | 90 | 9.3 | <0.01 | 165 | **78** | 0.39 | 92 | 8.6 | <0.01 |
| 12 | 5.84E + 06 | 6.37E + 05 | 0.98 | 90 | 220.9 | <0.01 | 3.42E + 05 | **1.11E + 04** | 0.3 | 90 | 152.7 | <0.01 | 1.00E + 04 | **2.85E + 03** | 0.13 | 83 | 14.8 | <0.01 |
| 13 | 6.75E + 13 | 1.02E + 14 | 1.00 | 90 | 9.14E + 07 | <0.01 | 3.05E + 08 | 9.66E + 07 | −0.01 | 88 | 0 | 0.98 | 8.37E + 09 | **9.40E + 07** | 0 | 84 | 408 | <0.01 |
| 14 | 1.62E + 12 | 1.61E + 12 | 0.98 | 90 | 0.1 | 0.93 | 1.80E + 04 | **1.02E + 04** | 0.29 | 90 | 6.1 | <0.01 | **1.02E + 09** | 1.56E + 09 | −0.02 | 91 | 4.2 | <0.01 |
| 15 | 1.49E + 09 | 2.30E + 09 | 0.80 | 90 | 7.1 | <0.01 | 5.24E + 08 | **4.42E + 07** | −0.01 | 90 | 55.9 | <0.01 | 8.25E + 08 | **4.12E + 08** | −0.01 | 92 | 7.2 | <0.01 |
| 16 | 9.94E + 06 | 4.16E + 08 | -0.01 | 90 | 198.5 | <0.01 | **1.04E + 08** | 1.05E + 09 | −0.02 | 89 | 46.9 | <0.01 | 2.18E + 09 | **6.07E + 07** | 0.94 | 82 | 487 | <0.01 |

Significantly smaller standard deviations of the participants' second-round estimates are in bold font.

Table 5. One-tailed Wilcoxon signed-rank sum tests comparing the absolute difference between first round and second round accuracy scores for best participants versus the absolute difference between first-round and second-round estimates for other participants for GI and GG conditions

| | *W* value (*N*) | *p* value |
|---|---|---|
| 1 | 2394 (141) | <**0.01** |
| 2 | 3447 (148) | <**0.01** |
| 3 | 2152 (136) | <**0.01** |
| 4 | 4151 (160) | <**0.01** |
| 5 | 3555 (136) | 0.36 |
| 6 | 9366 (160) | 1 |
| 7 | 7788 (149) | 1 |
| 8 | 1452 (161) | <**0.01** |
| 9 | 3696 (153) | <**0.01** |
| 10 | 5124 (145) | 0.65 |
| 11 | 3163 (154) | <**0.01** |
| 12 | 4153 (157) | <**0.01** |
| 13 | 1460 (155) | <**0.01** |
| 14 | 4634 (155) | **0.03** |
| 15 | 3142 (165) | <**0.01** |
| 16 | 3874 (152) | <**0.01** |

Statistically significant *p*-values are indicated in bold font.

questions where GG and GI participants were more accurate appear in bold font. Mann–Whitney *U* test statistics are also provided for each question. In both group conditions, accuracy scores were better than those in the II condition for 69% of the questions.[4] The benefits of independence (uncorrelated errors) were outweighed by the benefits of group interaction (information sharing).

### Social influence increases consensus

Pitman–Morgan tests allowed us to examine differences in variability of estimates in the first and second rounds of the group conditions.[5] Table 4 shows the results. Significantly smaller standard deviations of the participants' second-round estimates are in bold font. Although there was no consistent convergence on the second-round estimates in the II condition, talking in groups increased consensus on virtually all questions. In the GG condition, second-round estimate variability was reduced for 88% of the questions. In the GI condition, second-round estimate variability was reduced for 81% of the questions. Thus, our hypothesis about increased consensus after group discussion was supported.

### Does process loss occur within groups?

Research on groups often compares aggregate performance with that of the best-performing individual (Henry, 1995). Can groups become as "intelligent" as the wisest member? According to Parenté and Anderson-Parenté (1987), individuals who are less knowledgeable should be able to adjust their estimates through reflection on feedback and, in the process, get closer to the median of the group, whereas those who are more knowledgeable should maintain their estimates.

[4]The binomial probability of having 10 successes (assuming $p = 0.5$) is 0.12.
[5]Pitman–Morgan tests are used to compare the variance of correlated observations.



**DOI**: 10.1002/bdm

Table 6. One-tailed Wilcoxon rank-sum sign tests comparing the absolute difference between first-round and second-round accuracy scores for best participants versus the absolute difference between first-round and second-round median accuracy scores without the best participants for GI and GG conditions

| | *W* value (*N*) | *p* value |
|---|---|---|
| 1 | 136 (18) | 0.02 |
| 2 | 126 (19) | 0.11 |
| 3 | 164 (18) | <0.01 |
| 4 | 164 (20) | 0.01 |
| 5 | 124 (18) | 0.05 |
| 6 | 27 (20) | 0.99 |
| 7 | 33 (19) | 0.99 |
| 8 | 210 (20) | <0.01 |
| 9 | 169 (19) | <0.01 |
| 10 | 150 (19) | 0.01 |
| 11 | 191 (20) | <0.01 |
| 12 | 168 (19) | <0.01 |
| 13 | 189 (19) | <0.01 |
| 14 | 170 (20) | <0.01 |
| 15 | 208 (20) | <0.01 |
| 16 | 171 (19) | <0.01 |

Table 7. One-tailed Wilcoxon rank-sum sign tests comparing the absolute difference between first-round and second-round accuracy scores for worst participants versus the absolute difference between first-round and second-round estimates for other participants for GI and GG conditions

| | *W* value (*N*) | *p* value |
|---|---|---|
| 1 | 8993 (145) | <0.01 |
| 2 | 10607 (157) | <0.01 |
| 3 | 7445 (142) | <0.01 |
| 4 | 11283 (163) | <0.01 |
| 5 | 7995 (153) | <0.01 |
| 6 | 6625 (154) | <0.01 |
| 7 | 6203 (147) | <0.01 |
| 8 | 13366 (163) | <0.01 |
| 9 | 10694 (155) | <0.01 |
| 10 | 11691 (163) | <0.01 |
| 11 | 12485 (165) | <0.01 |
| 12 | 12204 (157) | <0.01 |
| 13 | 12560 (159) | <0.01 |
| 14 | 10608 (156) | <0.01 |
| 15 | 12109 (166) | <0.01 |
| 16 | 12031 (157) | <0.01 |

We tested this claim by comparing the difference between first-round and second-round estimates for the best individuals with those of others. To run this analysis, we calculated the difference between first-round and second-round estimates for all individuals. Then we picked the individuals with the best accuracy scores for first-round estimates in each group for each question and compared the difference between estimates in the two rounds with those shown of the other group members. If more than one person had the best accuracy score, we averaged the difference between first-round and second-round estimates over all best individuals in a given group, and compared this difference with the other differences. For each question, we used a one-tailed Wilcoxon signed-rank sum test to compare the difference between first and second rounds of best individuals with those of others.

Table 5 shows the test statistics for each question. The results showed that for 75% of the questions, best individuals showed less change between their first-round and second-round estimates than the others. Our results support Parenté and Anderson-Parenté's hypothesis that more knowledgeable participants will show less change in their estimates after group discussion.

Additionally, we ran the analysis by comparing the average group difference between the two rounds without the best individuals with the difference of the best individuals. Table 6 shows test statistics for each question. For 81% of the questions, the best individuals showed less change between their first-round and second-round estimates even when compared with the average group change.

Finally, tests comparing the difference of the worst participants with the difference of others showed that, in all questions, the worst individuals changed their estimates more than others. Table 7 shows the test statistics for each question. Less knowledgeable participants adjust their estimates more, which is in agreement with Parenté and Anderson-Parenté's hypothesis.

Even though the convergence on the median value might be beneficial, the hypothesis put forth by Parenté and Anderson-Parenté (1987) does not suggest how individuals might improve even further. A complementary theory was put forward by Laughlin and Ellis (1986) that asserts a group can reach the maximum level of performance (or for our purposes, accuracy) if the experts within a group can demonstrate their expertise clearly to the others in their group.

In the Delphi method, group discussion may help individuals realize their shortcomings and motivate them to make necessary adjustments. But there may still be flaws to this method. Bolger and Wright (2011) argued that remnants of identity might still be present through the expression of confidence ratings. Moreover, the Delphi method does not prevent people from egocentric discounting, where people give more weight to their own judgments regardless of their

Table 8. Comparisons between accuracy of best individual in Round 1 and group estimates in Round 2

| | GG condition | | GI condition | |
|---|---|---|---|---|
| | *W* value (*N*) | *p* value | *W* value (*N*) | *p* value |
| 1 | 0 (10) | 0.978 | 0 (10) | 0.99 |
| 2 | 0 (10) | 0.985 | 1 (9) | 0.99 |
| 3 | 9 (10) | 0.957 | 3 (9) | 0.97 |
| 4 | 0 (10) | 0.997 | 4 (9) | 0.99 |
| 5 | 2 (10) | 0.986 | 5 (9) | 0.96 |
| 6 | 0 (10) | 0.997 | 2 (10) | 0.99 |
| 7 | 0 (10) | 0.995 | 2 (9) | 0.99 |
| 8 | 3 (10) | 0.979 | 0 (10) | 0.99 |
| 9 | 0 (9) | 0.990 | 11 (10) | 0.87 |
| 10 | 0 (9) | 0.997 | 3 (9) | 0.99 |
| 11 | 11 (10) | 0.959 | 1 (10) | 0.99 |
| 12 | 7 (10) | 0.986 | 1 (10) | 0.95 |
| 13 | 9 (10) | 0.908 | 2 (8) | 0.97 |
| 14 | 0 (10) | 0.993 | 1 (10) | 0.99 |
| 15 | 4 (9) | 0.983 | 7 (9) | 0.95 |
| 16 | 35 (10) | 0.086 | 3 (9) | 0.99 |

Test statistics are Wilcoxon rank sums.



**DOI**: 10.1002/bdm

level of expertise or ensure the optimal level of information sharing, which might lead to ignorance of minority opinion even if it is more accurate. All these factors might be the underlying causes for the process loss observed in the Delphi method.

Process loss can be operationally defined as the difference between the best individual and group estimates. We examined process losses in Round 2 using one-tailed Wilcoxon signed-rank sum tests. Table 8 shows that, in both group conditions and for all questions, aggregate estimate in the second round was less accurate than the most accurate subject in the first round. Even though we have observed some process gain in groups as discussed earlier, groups were not using all of the available information, perhaps because participants did not know each other beforehand and were unfamiliar with each other's areas of expertise.

**Confidence did not change after social influence**

Lorenz et al. found that social influence resulted in greater confidence, despite no improvement in accuracy. We compared confidence ratings in Rounds 1 and 2 for the II, GG, and GI conditions and found no systematic differences. Table 9 shows median confidence ratings for all conditions. Questions for which confidence was significantly greater in the second round are marked in bold font, and the one-tailed Wilcoxon signed-rank sum test statistic is provided for each pair.

There was no change in confidence ratings of the II participants. When groups had individualized incentives (GI), confidence increased after discussion half the time (50% of questions), and in 75% of those questions, accuracy improved. With group incentives (GG), there were no systematic changes in confidence over rounds. Confidence increased in only two of the questions, and in only one of those questions did accuracy actually improve. In seven of eight questions, accuracy improved without any increase in confidence ratings. Thus, our hypothesis was not supported; there were no systematic effects of social influence on participants' confidence ratings.

**No difference between individual and group incentives**

Contrary to our predictions, accuracy did not improve with group incentives. Table 10 shows results from two-tailed Mann–Whitney *U* tests. Although GG participants were more accurate than GI participants in half of the questions, GI participants were more accurate than GG participants in the other half. There are many differences between our study and prior work showing that group incentives increase accuracy. In previous research, groups were dyads or triads. Participants interacted in person, and they received feedback after each question. In our study, groups consisted of 8–10 people. They interacted online, and they received no feedback until after they had completed all of the questions. Effects of incentives appear to be more complex than we initially anticipated.

**Capturing the process of group interaction**

We proposed a confidence-weighted model to describe the effects of group interaction in which second-round estimates assumed to be represented as a weighted average of an

Table 10. Effects of incentives on accuracy: median accuracy scores and Mann–Whitney *U* test results comparing GI and GG conditions' second-round accuracy scores

| | GI | GG | *U* value ($N_1$, $N_2$) | *p* value |
|---|---|---|---|---|
| 1 | 2.9 | 1.9 | 4074 (92, 94) | 0.49 |
| 2 | 8.6 | 13.6 | 4553 (92, 93) | 0.45 |
| 3 | **8** | 13.1 | 5465 (92, 94) | <0.01 |
| 4 | 7 | 5 | 3807 (92, 94) | 0.16 |
| 5 | **9** | 9 | 5182 (92, 93) | 0.01 |
| 6 | 33 | **32** | 3594 (92, 94) | 0.04 |
| 7 | 25 | 23 | 3717 (92, 93) | 0.12 |
| 8 | 0.8 | 0.8 | 4595 (92, 93) | 0.38 |
| 9 | 15 | 15 | 4313 (91, 93) | 0.82 |
| 10 | 17 | 17 | 3651 (91, 92) | 0.13 |
| 11 | 64 | 64 | 4137 (92, 94) | 0.61 |
| 12 | 1148 | 948 | 4172 (92, 94) | 0.68 |
| 13 | 7099 | 9901 | 4284 (92, 94) | 0.87 |
| 14 | **12 184** | 17 184 | 5167 (92, 94) | 0.01 |
| 15 | 3 672 441 | 3 679 191 | 4401 (92, 94) | 0.64 |
| 16 | 2 225 200 | 2 174 800 | 3793 (91, 93) | 0.15 |

Table 9. Median confidence ratings and Wilcoxon signed-rank sum statistics

| | II first | II second | *W* value (*N*) | *p* value | GG first | GG second | *W* value (*N*) | *p* value | GI first | GI second | *W* value (*N*) | *p* value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 40 (92) | 0.97 | 4 | 3.5 | 964 (61) | 0.45 | 4 | 3 | 1002 (67) | 0.81 |
| 2 | 2 | 2 | 182 (92) | 0.08 | 3 | 3 | 728 (60) | 0.92 | 3 | 4 | 1627 (67) | <**0.01** |
| 3 | 3 | 2 | 221 (92) | 0.48 | 4 | 4 | 1119 (71) | 0.82 | 4 | 4 | 721 (48) | 0.08 |
| 4 | 2 | 2 | 178 (92) | 0.82 | 3 | 3 | 929 (60) | 0.46 | 3 | 4 | 1380 (67) | 0.06 |
| 5 | 4 | 4 | 201 (92) | 0.39 | 4 | 4 | 692 (62) | 0.98 | 4 | 3 | 901 (67) | 0.94 |
| 6 | 2 | 2 | 138 (92) | 0.94 | 4 | 3 | 804 (62) | 0.89 | 4 | 4 | 1018 (60) | 0.22 |
| 7 | 3 | 2 | 156 (92) | 0.43 | 4 | 4 | 1324 (67) | 0.12 | 4 | 4 | 1346 (61) | <**0.01** |
| 8 | 2 | 2 | 385 (92) | 0.41 | 3 | 3 | 1496 (72) | 0.15 | 2 | 3 | 1524 (66) | <**0.01** |
| 9 | 2 | 2 | 196 (92) | 0.79 | 4 | 4 | 981 (64) | 0.66 | 3 | 4 | 1175 (60) | **0.03** |
| 10 | 2.5 | 3 | 217 (92) | 0.51 | 4 | 4 | 1107 (62) | 0.18 | 3 | 4 | 1097 (59) | 0.05 |
| 11 | 2 | 2 | 206 (92) | 0.57 | 3 | 3 | 962 (56) | 0.09 | 4 | 4 | 963 (61) | 0.45 |
| 12 | 2 | 2 | 294 (92) | <**0.01** | 3 | 3 | 1126 (69) | 0.69 | 3 | 4 | 1283 (59) | <**0.01** |
| 13 | 2 | 1.5 | 198 (92) | 0.28 | 3 | 4 | 1440 (70) | 0.12 | 3 | 4 | 1130 (55) | <**0.01** |
| 14 | 2 | 3 | 190 (92) | 0.75 | 3 | 3 | 1256 (63) | **0.04** | 3 | 3 | 1569 (70) | **0.02** |
| 15 | 1 | 1 | 182 (92) | 0.29 | 3 | 3 | 1347 (64) | **0.02** | 3 | 4 | 1683 (68) | <**0.01** |
| 16 | 2 | 2 | 417 (92) | 0.35 | 3.5 | 3 | 1151 (68) | 0.56 | 3 | 3 | 870 (56) | 0.27 |

Questions for which confidence was significantly greater in the second round are marked in bold font.



**DOI**: 10.1002/bdm

individual's first estimate and the group median of first estimates, with weights depending on the individual's confidence in his or her initial estimate. To see how well the model described the data, we computed predictions and calculated accuracy scores for predicted estimates by taking the absolute value of the difference between $\hat{e}_{2ij}$ and true values.

We compared the accuracy of first-round scores with the accuracy of second-round predictions to see if the model predicted greater accuracy on the second round. Questions with statistically significant increases in predicted second-round accuracy are shown in bold font in Table 11. Predicted second-round estimates were more accurate than the first-round estimates in 94% of questions. The model was able to describe the data and capture the beneficial effects of social influence.

Next, we examined whether the model could account for increased consensus on second rounds. Pitman–Morgan tests were used to examine differences in variability of first-round and predicted second-round estimates pooled over both group conditions. Results appear in Table 12. Relative to first-round estimates, predicted second-round estimates revealed less variability in all of the questions. The model was able to capture another important effect of social interaction.

Table 11. One-tailed Wilcoxon rank-sum sign tests comparing the absolute difference between first-round and second-round accuracy scores versus the absolute difference between first-round and predicted second-round estimates

| | *W* value (*N*) | *p* value |
|---|---|---|
| 1 | 6593.5 (186) | 0.08 |
| 2 | 4768.5 (175) | **0.002** |
| 3 | 3118.5 (177) | **<0.001** |
| 4 | 2337.5 (175) | **<0.001** |
| 5 | 4142.5 (178) | **<0.001** |
| 6 | 4339.5 (184) | **<0.001** |
| 7 | 3998.5 (183) | **<0.001** |
| 8 | 2554 (186) | **<0.001** |
| 9 | 4212 (177) | **0.005** |
| 10 | 1852.5 (163) | **<0.001** |
| 11 | 3992 (185) | **<0.001** |
| 12 | 3542 (183) | **<0.001** |
| 13 | 2487.5 (160) | **<0.001** |
| 14 | 3553.5 (186) | **<0.001** |
| 15 | 1601 (177) | **<0.001** |
| 16 | 3162 (175) | **<0.001** |

## A closer look at online group interactions

Having found benefits of group interaction, we looked deeper into possible causes. The following *post hoc* analyses provide some answers. Three raters who were not involved in the study were recruited to code group discussions. They were given 11 different categories of comment types and were asked to code each line of discussion. The first coding was a "1" if the specific line met the description of any given category, and "0" if not. A line could meet the description of multiple categories. Categories were as follows: (1) providing information that others may not know; (2) providing an estimate or range; (3) reasserting an estimate; (4) changing an estimate; (5) self-identifying as an expert; (6) searching for an expert; (7) claiming outside knowledge other than being an expert; (8) contradicting someone else's claims; (9) questioning someone's claims; (10) agreeing with someone else's claims; and (11) making irrelevant comments or asking irrelevant questions.

After the raters finished coding, we looked at a measure of inter-rater agreement of multivariate observations known as the iota coefficient in the R package *irr* (Gamer, Lemon, Fellows, & Singh, 2012). The iota coefficient was 0.4, suggesting a fair agreement among our raters. We also examined Fleiss' kappa as an index of inter-rater agreement among our three raters using kappam.fleiss() function in the R package *irr* (Gamer et al., 2012). Values of kappa for different code categories are shown in Table 13. Because we had only fair agreement among our raters and a closer look at the frequencies of using a category showed great variance among our raters, if any of our raters coded any category as 1 for any given line, we assigned those lines to category 1.

To find out whether comments could predict group accuracy, we examined the correlation of the comment categories with each other. Three pairs of code categories had very high correlations: $\phi = 0.85$ for category 2, providing an estimate or range, and category 3, reasserting an estimate; $\phi = 0.64$ for category 1, providing information that others may not know, and category 7, claiming outside knowledge other

Table 12. Estimate standard deviations and Pitman–Morgan test statistics for participants' first-round and predicted second-round estimates

| | First round | Predicted second round | *r* | *df* | *t* | *p* |
|---|---|---|---|---|---|---|
| 1 | 690 940 600 | 249 076 300 | 0.97 | 184 | 72 | <0.001 |
| 2 | 798 | 158 | 0.75 | 174 | 48 | <0.001 |
| 3 | 246 479 | 73 893 | 1.00 | 175 | 689 | <0.001 |
| 4 | 1 508 404 000 | 188 583 500 | 1.00 | 174 | 762 | <0.001 |
| 5 | 749 531 700 000 000 000 000 000 | 2 163 882 000 | −0.01 | 176 | 2 297 718 000 000 000 | <0.001 |
| 6 | 725 839 700 | 285 915 200 | 0.71 | 183 | 20 | <0.001 |
| 7 | 2197 | 250 | 0.98 | 184 | 329 | <0.001 |
| 8 | 176 | 80 | 0.73 | 184 | 17 | <0.001 |
| 9 | 14 | 8 | 0.82 | 176 | 13 | <0.001 |
| 10 | 6 | 3 | 0.84 | 161 | 16 | <0.001 |
| 11 | 16 | 3 | 0.68 | 184 | 42 | <0.001 |
| 12 | 11 | 6 | 0.85 | 181 | 15 | <0.001 |
| 13 | 19 | 12 | 0.87 | 158 | 12 | <0.001 |
| 14 | 13 | 7 | 0.83 | 184 | 16 | <0.001 |
| 15 | 14 | 7 | 0.86 | 176 | 17 | <0.001 |
| 16 | 17 | 10 | 0.85 | 174 | 15 | <0.001 |



**DOI**: 10.1002/bdm

Table 13. Fleiss' kappa values for chat code categories

| Coding categories | $\kappa$ |
|---|---|
| 1. Providing information that others may not know | 0.55 |
| 2. Providing an estimate or range | 0.33 |
| 3. Reasserting an estimate | 0.44 |
| 4. Changing an estimate | −0.02 |
| 5. Self-identifying as an expert | 0.36 |
| 6. Searching for an expert | 0.41 |
| 7. Claiming outside knowledge other than being an expert | 0.15 |
| 8. Contradicting someone else's claims | 0.51 |
| 9. Questioning someone else's claims | 0.27 |
| 10. Agreeing with someone else's claims | 0.59 |
| 11. Making irrelevant comments/asking irrelevant questions | 0.33 |

than being an expert; and $\phi = 0.41$ for category 6, searching for an expert, and category 9, questioning someone's claims. We then combined each pair into a single value as Est, Info, and Ask, respectively; we coded these values as 1 if either of the components was present. Although (3) and (4) were correlated ($\phi = 0.41$), we did not create a new variable by combining these two categories, because, on theoretical grounds, we expected "changing an estimate" to be particularly important. We thus reduced the codes to predictors. We then examined the main dependent variable, accuracy, and found that its distribution was extremely skewed. We added 1 and took the logarithm, thereby removing most of the skew. We then asked whether the seven categories predicted the logarithm of second-round group median accuracy score. To assess this, we used the lmer() function in the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2014), and we report *p*-values resulting from parametric bootstrapping, using the PBmodcomp() function in the R package *pbkrtest* (Højsgaard, 2013). For all tests, the logarithm of second-round group median accuracy score was the dependent measure; the coding categories changing an estimate, self-identifying as an expert, contradicting someone else's claims, agreeing with someone else's claims, Est (providing an estimate or reasserting an estimate), Info (providing information that others may not know or claiming outside knowledge other than being an expert), and Ask (searching for an expert or questioning someone's claims) were the fixed-effect terms; and groups and questions were the random-effect terms.[6]

Regressing the logarithm of group median accuracy scores on all the seven coding variables mentioned earlier yielded a significant effect ($X^2 = 5.22$, $p = 0.0120$). $R^2$ for the correlation between the model predictions and the data was 0.029. To examine the sources of this effect, we tested each predictor one at a time. Three of the predictors yielded significant results (uncorrected for multiple tests): changing an estimate, Est (providing an estimate or reasserting an estimate), and Ask (searching for an expert or questioning someone's claims). Our accuracy measure is actually a measure of error, so changing estimates and making more estimates were beneficial to accuracy, whereas asking for information was negatively related to accuracy, perhaps because this occurred when participants felt correctly that they were ignorant.

[6]Because of programming error, we had data only from 15 groups rather than the 20 groups we tested.

Our chat analysis results showed a beneficial effect of changing estimates. Groups performed better when their members were more open-minded about receiving input. This result is consistent with correlations between actively open-minded thinking and accuracy in other studies (Haran, Ritov, & Mellers, 2013; Mellers, Stone, et al., 2014). Our results also revealed a positive effect of providing more estimates, also consistent with previous findings that active forecasters do better in prediction tasks (Mellers, Stone, et al., 2014).

## CONCLUSION

We asked participants to make a series of estimates of unknown quantities, such as the height of the tallest man-made building in the world. Aggregate estimates from interacting groups were more accurate than aggregate estimates of individuals working alone. The benefits of information sharing within groups outweighed the costs of greater dependencies in errors. Benefits were apparent in both within-subject and between-subject comparisons.

Our results may seem at odds with those of Lorenz et al. (2011).[7] Lorenz et al. (2011) concluded that group interaction provided no improvement in accuracy over individuals working alone in a similar task, despite an increase in confidence and consensus. However, Lorenz et al. (2011) used a different definition of accuracy (the log of the ratio of the estimate relative to truth) and different statistical tests to assess the data (i.e., Kolmogorov–Smirnov tests). Kolmogorov–Smirnov tests are advantageous because they are based on cumulative density functions for comparing samples from different populations and do not require assumptions of normality. The Kolmogorov–Smirnov test, however, does not accommodate the within-subject nature of the design of these studies. The Wilcoxon rank sum tests (for within-subject comparisons) and Mann–Whitney *U* tests (for between-subject comparisons) that we used are more powerful.

Differences between our results and those of Lorenz et al. could be due to different definitions of accuracy or different statistical tests. We reanalyzed the data of Lorenz et al. using our definition of accuracy (absolute value of the difference between estimate and truth) and our tests (Mann–Whitney *U* and Wilcoxon rank sum). The results of Lorenz et al. were perfectly consistent with ours; both experiments showed that social influence had beneficial effects. Our analyses of Lorenz et al. data can be found at http://finzi.psych.upenn.edu/~baron/burcu/.

We also applied the definition of accuracy of Lorenz et al to our data to see if the conclusions would change. We used a logarithmic transformation of the ratio of the estimate to truth. Conclusions about our data based on their definition of accuracy were virtually the same as conclusions about their data based on their definition of accuracy. There

[7]A letter by Farrell (2011) also challenges the claims of Lorenz et al. of how social influence undermines the wisdom of crowds effect by looking at the monetary rewards participants would earn before and after the information exchange and shows that the decrease in variance is beneficial.



**DOI**: 10.1002/bdm

appears to be no conflict between our data and those of Lorenz et al.

Other studies have found that groups can be superior to individuals working alone. In a mock trial jury, group discussion yielded better recall of information (Vollrath, Sheppard, Hinsz, & Davis, 1989). In a real-world forecasting tournament of geopolitical events, groups were more accurate than individuals working alone (Mellers, Ungar, et al., 2014). Groups shared information and rationales. They motivated team members, and they corrected obvious errors in each other's forecasts. Understanding how to structure groups to boost performance even more requires greater theoretical and empirical insight. But for now, the results are in; social effects bolster and strengthen accuracy when numerical estimations are at stake.

## ACKNOWLEDGEMENTS

The authors thank Jierui Song, Young H. Lee, Janna Rapoport, Nick Rohrbaugh, Nadia Ogene, Katsiaryna Malykhina, Leah Fang, and The Wharton Behavioral Lab for valuable assistance. This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

## REFERENCES

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal and Social Psychology*, *67*, 422–436.

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), Advances in experimental social psychology, *2*, (pp. 267–296). San Diego, CA: Academic Press.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), Groups, leadership, and men. Pittsburgh, PA: Carnegie Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1.-7. https://github.com/lme4/lme4/http://lme4.r-forge.r-project.org/ (accessed on April 2013).

Beersma, B., Hollenbeck, J. R., Humphrey, S. E., Moon, H., Conlon, D. E., & Ilgen, D. R. (2003) Cooperation, competition, and team performance: Toward a contingency approach. *Academy of Management Journal*, *46*, 572–590.

Best, R. J. (1974). An experiment in Delphi estimation in marketing decision making. *Journal of Marketing Research*, *11*, 448–452.

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting & Social Change*, *78*, 1500–1513.

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting, *Journal of Management Accounting Research*, *12*, 19–64.

Branson, L., Steele, N. L., & Sung, C. (2010). When two heads are worse than one: Impact of group style and information type on performance evaluation, *Journal of Business and Behavioral Sciences*, *22*(1), 75–84.

Brown, B. B. (1968). Delphi process: A methodology used for the elicitation of opinions of experts: An earlier paper published by RAND (document no: P-3925, 1968).

Cooper, R. S. (1991). Information processing in the judge-adviser system of group decision-making. *Unpublished master's thesis*. University of Illinois, Urbana–Champaign.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9*(3), 458–467.

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*, 22–49.

DeMatteo, J. S., Eby, L. T., & Sundstrom, E. (1998). Team-based rewards: Current empirical evidence and directions for future research. *Research in Organizational Behavior*, *20*, 141–183.

Deutsch, M. (1949). A theory of cooperation and competition. *Human Relations*, *2*, 129–152.

Farr, J. L. (1976). Incentive schedules, productivity, and satisfaction in work groups: A laboratory study. *Organizational Behavior and Human Performance*, *17*, 159–170.

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *PNAS*, *108*: E625.

Fraidin, S. N. (2004). When is one head better than two? Interdependent information in group decision making. *Organizational Behavior and Human Decision Processes*, *93*, 102–113.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). Irr: Various coefficients of interrater reliability and agreement. R package version 0.84.

Gigone, D., & Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*, *65*, 959–974.

Halekoh, U., & Højsgaard, S. (2013). pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison. R package version 0.3-8. http://CRAN.R-project.org/package=pbkrtest (accessed on May 2014).

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*(3), 188–201.

Heneman, R. L., & von Hippel, C. (1995). Balancing group and individual rewards: Rewarding individual contributions to the team. *Compensation and Benefits Review*, *274*, 63–68.

Henry, R. (1995). Improving group judgment accuracy: Information sharing and determining the best member. *Organizational Behavior and Human Decision Processes*, *62* (8), 190–197.

Hertel, G., Kerr, N. L., & Messé, L. A. (2000). Motivation gains in performance groups: Paradigmatic and theoretical developments on the Kohler effect. *Journal of Personality and Social Psychology*, *79*, 580–601.

Janis, I. (1982). Group-think, Boston, MA: Houghton-Mifflin.

Kaplan, A., Skogstad, A. L., & Girshick M. (1949). The prediction of social technological events, Rand Corp. P-93.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655.

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*, 77–124.

Landeta, J. (2006) Current validity of the Delphi method in social sciences, *Technological Forecasting & Social Change*, *73*, 467–482.

Larson, J. R., Jr., Foster-Fishman, P. G., & Keys, C. B. (1994). Discussion of shared and unshared information in decision-making groups. *Journal of Personality and Social Psychology*, *67*, 446–461.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *The Journal of Personality and Social Psychology*, *37*, 822–832.





**DOI**: 10.1002/bdm

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(2), 177–189.
Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology*, *90*, 644–651.
Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology*, *41*, 585–634.
Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect, *PNAS*, *108*(22), 9020–9025.
McGrath, J. (1984). Groups: Interaction and performance. Englewood Cliffs, NJ: Prentice Hall.
Mellers, B. A., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2014). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics, manuscript under review. Philadelphia, PA: University of Pennsylvania.
Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gürçay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S., Murray, T. & Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.
Michaelson, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group decision making. *Journal of Applied Psychology*, *72*, 834–839.
Parenté, F. J., & Anderson-Parenté, J. K. (1987). Delphi inquiry systems. In G. Wright and P. Ayton (Eds.), Judgmental forecasting. Chichester: Wiley.
Phillips, J. M. (1999). Antecedents of leader utilization of staff input in decision-making teams. *Organizational Behavior and Human Decision Processes*, *77*(3), 215–242.
Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*, 73–89.
Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting*, *15*, 353–375.
Rowe, G., Wright, G., & McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. *Technological Forecasting & Social Change*, *72*, 377–399.
Sackman, H. (1974). Delphi assessment: Expert opinion, forecasting and group process. *R-1283-PR*.
Sherif, M. (1936) The psychology of social norms. New York: Harper.
Stasser, G., Taylor, L. A., & Hanna, C. (1989). Information sampling in structured and unstructured discussions of three- and six-person groups. *Journal of Personality and Social Psychology*, *57*, 67–78.
Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology*, *53*, 81–93.
Surowiecki, J. (2004). The wisdom of crowds. New York: Doubleday.
Tindale, R. S., & Larson, J. R. (1992). Assembly bonus effect or typical group performance? A comment of Michaelsen, Watson, and Black (1989). *Journal of Applied Psychology*, *77*, 102–105.
Tindale, R. S., Smith, C. M., Thomas, L. S., Filkins, J., & Sheffey, S. (1996). Shared representations and asymmetric social influence processes in small groups. In E. Witte, J. H. David (Ed.), Understanding group behavior: Consensual action by small groups (*1*, pp. 81–103), Mahwah, NJ: Erlbaum.
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.
Vollrath, D. A., Sheppard, B. H., Hinsz, V. B., & Davis, J. H. (1989). Memory performance by decision-making groups and individuals, *Organizational Behavior and Human Decision Processes*, *43*, 289–300.
Welbourne, T. M., & Gomez Mejia, L. R. (1995). Gainsharing—A critical-review and a future-research agenda. *Journal of Management*, *21*, 559–609.
Wittenbaum, G. M., & Park, E. S. (2001). The collective preference for shared information. *Current Directions in Psychological Science*, *10*, 70–73.
Wittenbaum, G. M., & Stasser, G. (1996). Management of information in small groups. In J. L. Nye, A. M. Brower (Eds.), What's social about social cognition (pp. 967–978). Thousand Oaks, CA: Sage.

*Authors' biographies*:

**Burcu Gürçay** is a Doctoral Candidate in Psychology at the University of Pennsylvania. Her research focuses on moral judgment models and forecasting.

**Barbara A. Mellers** is Professor of Psychology and Marketing at the University of Pennsylvania. She studies factors that influence human judgments and decisions, including emotions, contextual effects, and response modes. She has developed models of fairness and cooperation in economic games. Her current research examines intuitive forecasts of geopolitical events.

**Jonathan Baron** is Professor of Psychology at the University of Pennsylvania.

*Authors' addresses*:

**Burcu Gürçay**, Department of Psychology, University of Pennsylvania, Philadelphia, PA USA.
**Barbara A. Mellers**, Department of Psychology, University of Pennsylvania, Philadelphia, PA USA.
**Jonathan Baron**, Department of Psychology, University of Pennsylvania, Philadelphia, PA USA.



**DOI**: 10.1002/bdm