Information-Pooling in Wisdom of Crowds Judgments

Reid Hastie, John G. Burrows, Chen Xia
Booth Graduate School of Business, University of Chicago
(contact: reid.hastie@chicagobooth.edu)

We report on an empirical and theoretical exploration of the conditions under which information-pooling will increase the accuracy of group judgments. For the most part, wisdom-of-crowds tests of group judgment accuracy have not found dramatic advantages for groups that share information among members before rendering individual or consensus judgments. Yet, idealized examples imply that there should be conditions under which information-pooling will increase accuracy. Our answer to this puzzle is that there are some, but not all judgment problems where information-pooling improves performance. We attempt to (i) demonstrate empirically a clear information-pooling advantage for certain judgment problems; and (ii) spell-out conditions that could be used to identify those problems in advance of performance to know when to promote information-pooling, and when information-pooling will not be productive.

Information-Pooling and Wisdom of Crowds Judgments

1. Human beings have a distinctively complex social existence. Compared to other animals, we require extensive social contact to develop normally. We have an unusually elaborate collection of communication methods, capped by the achievement of oral and written languages. And, most of us interact frequently and flexibly with many more members of our own families and species than any other creature.

2. Many of our greatest accomplishments are achieved via language, including extensive culture transmission (e.g., the explicit and tacit knowledge to create skyscrapers and AK-47s) and systematic social decision-making rituals (e.g., hierarchical principal-agent structures and ad hoc social networks). In fact, some social philosophers have advocated deliberative democracies as the ultimate social decision process.

3. The present essay is a conceptual and empirical exploration of a narrowly-focused social judgment process, the use of "Wisdom Of Crowds" mechanisms to make simple judgments and decisions. In a typical example of such an method, a small group of experts, perhaps business consultants or intelligence analysts, is tasked with making a sharply-focused factual estimate such as the price of a barrel of oil next month, or whether Greece will remain within the European political and economic union for another year. Three types of mechanisms are employed to make these social judgments: (i) a face-to-face discussion followed by a formal or informal consensus statement (e.g., a vote); (ii) a

structured social aggregation device such as a Delphi Method, a Nominal Group Method, or an Information (Prediction) Market; (iii) a no-discussion voting or statistical aggregation method. One manner in which these mechanisms can be distinguished is according to the degree of information-pooling included in each method. In free-wheeling face-to-face discussions there is considerable information-pooling, sharing solutions, opinions, and reasons or evidence for and against solutions. At the other end of this continuum are voting and statistical aggregation mechanisms where maximum independence and zero information-pooling is often advocated. In-between are mechanisms that may regulate information-pooling or simply ignore it without providing rules or provisions for sharing (or not).

There is long history of scientific interest in comparisons between individual versus group judgments and decisions (see Lorge, Fox, Davitz, & Brenner, 1958; Einhorn, Hogarth, & Klempner, 1977; Hill, 1982; Hastie, 1986; Gigone & Hastie, 1997; and Tindale & Kluwe, 2016, for comprehensive reviews of behavioral research). There is a smattering of behavioral studies focused on the question of which type of aggregation process, social versus statistical, produces the best performance, usually measured by estimation accuracy (Fischer, 1981; Graefe & Armstrong, 2010; Gustafson, Shukla, Delbecq, & Walster, 1973; Van de Ven & Delbecq, 1974).

4. Brief note on the scope of this essay: Political analysis, decisions, and discourse usually involve a mix of significant belief and preference issues. We argue about different preferences (goals such as maximizing patient quality-adjusted-life-years or minimizing expenditures to maintain lives beyond 70 years) for a medical policy and we argue about the facts relevant to achieving those goals (beliefs about which therapies will be effective and how much they will cost); we argue about our preferences for same-sex civil unions (e.g., the morality of homosexual and lesbian relationships), and we argue about the factual consequences of one policy versus another (e.g., the mental health of children raised by same-sex couples); ditto for every political topic from capital punishment to differential taxation of different income levels. To keep our discussion focused, we will restrict our consideration to judgment and decision tasks in which all participants share a common incentive structure to achieve maximal factual accuracy; where discussion is about what to believe, not what to value or prefer. At the end of the essay we will make a few remarks about the similarities and differences of persuasion on beliefs versus preferences.

5. Returning to our focus on groups attempting to maximize accuracy in discerning or forecasting states of the world, there is disagreement about the value of independence versus dependence across individual decision-

makers or judges. One view, associated with a statistical analogy, and often with statistical aggregation algorithms, favors independence and zero information pooling. Lorenz et al (2011). In the extreme, proponents of the statistical analogy advocate polling large, diverse, independent collections of informed judges for best solutions (usually estimates or forecasts), with the summary judgment derived by calculating a simple average (Surowiecki, 2004; some favor medians, Galton, 1907; while others favor geometric means, Lorenz et al, 2011), or the outcome of a majority-plurality vote (Hastie & Kameda, 2005).

The Wisdom of Crowds mechanism that increases accuracy in collective judgments is error-damping (statistical principles of sampling and numerical aggregation; or principles of voting, such as the Condorcet Jury Theorem).

The contrasting view, favors information-pooling, discussion, and reasoned deliberation before selecting or inferring a summary judgment; perhaps inspired by the analogy to a logical inference engine (Gurcay, Mellers, & Baron, 2014; Cohen, 1989; Guttman & Thompson, 2004). One of the rationales for the deliberative approach is the thought experiment suggested by the Hidden Profiles Paradigm for small group research (Stasser & Titus, 1985, 2003). The Hidden Profiles research task involves a judgment or inference problem in which a large number of factual evidential propositions imply one unique solution, usually according to a non-linear combination function. When the paradigm is applied in small group research the evidence is distributed across members of a team, such that no individual member has sufficient information to derive the correct solution implied by the complete set of propositions. In its most diabolical form, the propositions are distributed so that each individual team member has a subset that implies an incorrect solution. Given this initial information distribution state, the only way the group can find the correct solution is to pool the information during some form of deliberation so that at the end of discussion, the full set of propositions is available to members through that information-pooling process. The Hidden Profiles thought experiment implies that there are some judgment problems that <u>cannot</u> be solved by simply aggregating individual solutions by averaging or voting (as in the statistical approach). The correct solution can only be reliably attained through effective pooling of the unshared evidence. The most obvious method to promote the necessary information-pooling is some kind of vigorous verbal deliberation.

Here the accuracy-increasing mechanism is logical reasoning, with information-pooling providing a more complete set of premises from which to derive reasoned conclusions. And, the involvement of more reasonable (if not rational) agents providing more inferential and error-checking computational power.

6. There are a couple of other accuracy-enhancing social mechanisms, but we will concentrate on these two in this essay: error-damping and information-pooling. To be more complete, we should mention some of the other mechanisms. One mechanism that has received extensive empirical study is a “Truth-Wins, Solution Demonstrability” mechanism. When making some types of judgments (e.g., eureka brain-teaser puzzles, some mathematical calculation problems), when one member of a group solves the problem, the solution is self-verifying. This condition has been labeled “demonstrability.” So, for example, if one member of a group solves a “trick” brain-teaser such as one of the items on the popular Cognitive Reflection Test* or the economic Beauty Contest Game (Kocher & Sutter, 2005), the other members of the group are very likely to recognize the correct solution (and the solver is able to demonstrate the solution by describing the correct calculation). [FOOTNOTE: Frederick, 2005: “A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?” Common Answer: 10¢; Correct Answer: 5¢] A similar process occurs when the answer to a question depends on a clear area of expertise, and less-expert members of the group defer to the most expert member. [FOOTNOTE: “How many students are enrolled at the University of California at Irvine?” The group defers to the professor from Irvine (approx. 30,000). “What is the freezing point of ethyl alcohol?” The group defers to a chemist (–114°C).] And, finally, under some (not perfectly known) conditions, working in a team may increase effort and thus increase accuracy on problems that require extended efforts to solve.

7. If we restrict our attention to the error-damping and information-pooling mechanisms, the implication of our discussion thus far is that judgment problems that fit the Hidden Profiles, information distribution pattern will be the types of problems where discussion can have its biggest effects. Problems where everyone shares almost all the relevant evidence will not provide conditions where discussion can increase accuracy. Some examples of such problems might be estimating the number of beans in a mason jar (everyone has the same visual evidence) or estimating the height of the Sears Tower (almost everyone begins with the assumption the tower is approximately 100 stories tall). Another type of problem where discussion is unlikely to increase accuracy is one where no one has much relevant evidence (e.g., a sample of college students is asked to estimate the population density of Switzerland in 2006).

Thus, if we want to observe discussion, information-pooling effects on group judgment accuracy, we need to look for problems where discussion has a chance to contribute to accuracy, perhaps by creating a larger shared data base of relevant evidence. This insight is the inspiration for a small research program seeking judgment problems that produce higher accuracy, following discussion, compared to following solution sharing or following no communication. We will not summarize an initial empirical study designed to identify judgment problems that fit the Hidden Profiles pattern and that will demonstrate increased accuracy following discussion.

An Experimental Study of the Effects of Discussion on Small Group Accuracy

8. Empirical Study: Basic Procedure. Approximately 300 University of Chicago college students participated in a 40-minute experimental session making judgments of 8 objective quantities (and 2 qualitative facts). Eight of the 10 questions were designed to fit the Hidden Profiles pattern. Relying on intuitions, we produced 8 questions where we expected different randomly combined group members to possess substantial amounts of unshared information before attempting to answer the questions. Then, if discussion was effective, members would acquire relevant information, beyond their initial belief sets, and this additional information learned from the discussion, would improve individual and group accuracy.

To provide some controls on the social process, the experiment assigned participants to 3-member teams, and each team performed the judgment task (for the 10 questions) according to one Group Process Condition. Participants were incentivized for individual accuracy, by 3 $100 prizes, each paid to the most accurate individual in each of the 3 Group Process conditions.

The control condition was labeled “Solo”, where the 3 team members did not interact at all while they made 3 judgments for each question (although they were present in the laboratory at the same time, working independently at separate computer terminals). Note the Solo control condition controls for discussion or information-pooling [there was none] while still requiring the team members to make 3 judgments in a sequence. We expected to see some belief revision across the 3 trials on each question, as the participant reconsidered his or her answer on the previous round. Although we did not expect to observe increases in accuracy, or convergence between answers from different members of the non-interacting group.

The Judgment-Feedback Group condition involved making 3 judgments, and after each judgment each member of the team learned the judgments rendered by all 3 team members. Here, we expect there will be some dependence of the judgments of each member on the answers of the other team members. In fact, we encouraged them, in the general instructions for the task to, “Pay careful attention to the judgments of the other members,” and, “To think about making your next judgment within the range of judgments from the previous round.” Here we expected convergence among team members, and some increase in accuracy, especially as participants who initially rendered extreme, outlier answers would be likely to adjust dramatically after seeing more reasonable answers from the other members of their teams.

Finally, in the Discussion-Feedback Group condition, team members were instructed to communicate the reasons for their judgments on each trial, along with the judgments. ““What insight(s) helped you arrive at your answer for the this question? Remember we are looking for insights that you believe are likely to be UNshared by other participants answering the same question – insights that you had, that were relevant and significant in how you answered the question, but which are likely to have NOT been inferred or noticed by the other participants answering the same question. In other words, what are one or two of your UNIQUE (or at least unusual) INSIGHTS?” Here we expected to see the most convergence towards a common answer, and we hypothesized there would be increased accuracy as discussion produced more informed members through information-pooling on questions where we expected there would be a substantial amount of unshared, but relevant information.

The experimental design was defined on a sample of 300 participants, randomly assigned to 3-member teams in one of 3 Group Process conditions (Solo, Judgment-Feedback, Discussion-Feedback). Thus, there were approximately 100 participants and 30+ teams in each Group condition. Each team made 3 judgments in sequence with feedback on other member judgments (Judgment and Discussion) and relevant information (Discussion only), or no feedback of any kind (Solo). And, finally, teams made the 3 judgments on (up to) 10 Questions (n.b., some questions were added to the protocol part-way through the experiment, so Questions 9 and 10 were only answered by approximately 200 participants).

9. <u>Empirical Study: Test Questions</u>. The heart of this empirical project is the sample of judgment questions used to test for the effects of discussion on group judgment accuracy. As noted, at this point in the project, we relied on our intuitions to create judgment questions where we expected that there might be substantial unshared information before discussion occurred; in other words, questions where team members would have something to talk about. Here is a summary of the rationales behind our intuitions (see also the italicized comments in Table 1, below).

The types of questions for which we expect discussion will enhance accuracy are usually multiple-step problems, where the solution is not derived from a unitary, “eureka” insight, or from knowing one relevant item of evidence. In several cases we relied on the geographic diversity of the University of Chicago student body to increase the likelihood that a random sample of 3 members of a team would represent substantial samples of unshared pre-discussion information. In a few cases, we thought that deriving an answer would require several logical steps and that discussion could provide an additional error-checking, error-correction process.

Part-way through the experiment, we realized it would be important to include some questions where we believed (again based on intuitions) judgments would not be enhanced through discussion (see Questions 9 and 10 in Table 1).

TABLE 1. Questions 1 through 10 (excluding Raven's Progessive Matrices IQ Test, Question 7)

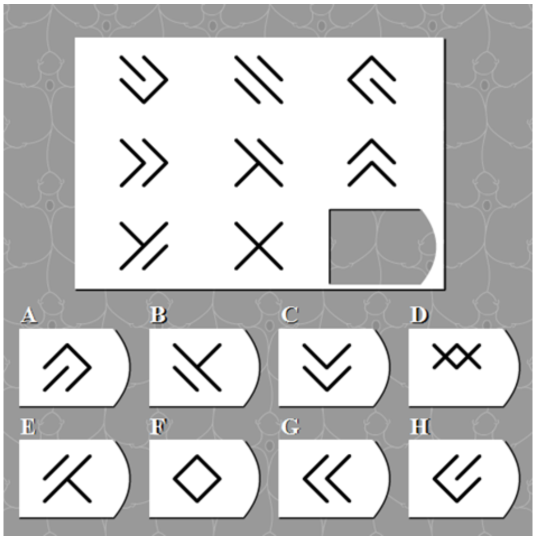
#	Question	Ans wer	Source/*Rationale*
1	How tall is the highest man-made building in feet?	2722	The world's tallest man-made structure (which includes “buildings”) is the 829.8 m (2,722 ft) tall Burj Khalifa in Dubai, United Arab Emirates. *Our thought here was that the question involved 2 parts – Which building is tallest? How tall is that building? So, that discussion might increase the relevant evidence for team members who did not know the answer to both parts.*
2	What is the total area of the contiguous U.S. in square miles? The contiguous United States is the 48 adjoining U.S. states that are south of Canada and north of Mexico, plus the District of Columbia, but excluding the states of Alaska and Hawaii, and all off-shore U.S. territories and possessions.	3,119,885	Together, the 48 contiguous states and Washington, D.C., occupy a combined area of 3,119,884.69 square miles, which is 1.58% of the total surface area of the Earth. *Here again, there are several parts to an answer (e.g., geographical locations) or that most effective solutions would require several steps to reach a conclusion (e.g., the Continental US is a*

			crude rectangle, the height would be …")
3	As of Financial Year 2012, how many Starbucks stores were there in the U.S.?	10,924	Source: http://www.statisticbrain.com/starbucks-company-statistics/ *The rationale is less compelling here, but again we thought this was a multi-part estimation problem*
4	The Chicago 'L' serves the city of Chicago and seven of its surrounding suburbs and is operated by the Chicago Transit Authority (CTA). As of 2013, how many stations were there spread across its 8 operating lines?	145	Source: http://en.wikipedia.org/wiki/Chicago_'L'#cite_note-CTA_facts-1 *This is the closest to an ideal paradigmatic question: Given that Chicago students would have experience with different parts of the L system, they would naturally have different pieces of the answer; discussion should increase everyone's knowledge base, and accuracy.*
5	According to data from the 2010 U.S. Census's American Community Survey, which of the following U.S. metropolitan areas has the largest percentage of foreign-born people?	Miami-Fort Lauderdale-Pompano Beach, FL	The other choices from highest to lowest were: San Jose-Sunnyvale-Santa Clara, CA; Los Angeles-Long Beach-Santa Ana, CA; New York-Northern New Jersey-Long Island, NY-NJ-PA; Chicago-Naperville-Joliet, IL-IN-WI *The thought here was that Chicago students would be likely to represent different geographic areas and could contribute to a more fully-informed judgment through*

			discussion.
6	What is the shortest total driving time (in whole hours, assuming average speed of 60 miles an hour and no stoppages) for a car journey beginning in Casper, Wyoming and ending in Key West, Florida, with stops on route in Indianapolis, Phoenix, Toledo, Louisville, and Trenton. Note: the stops can be completed in any order and assume no adverse weather conditions.	83 hours	According to Google Maps, the distance between Casper and Key West (stopping in order in Phoenix, Louisville, Indianapolis, Toledo, and Trenton) is 4953 miles. Question stated average speed of 60 miles per hour, so answer is 4953/60 = 83 hours. *Here we thought the problem is complex, with many parts, and Chicago students would be expert on different geographic locations*
8	What percentage of Americans has a pet?	62%	The Humane Society US suggests pet ownership in the U.S. has more than tripled from the 1970s, when approximately 67 million households had pets, to 2012, when there were 164 million owned pets. In other words, in 2012, 62 percent of American households included at least one pet.http://www.humanesociety.org/issues/pet_overpopulation/facts/pet_ownership_statistics.html *The rationale is less compelling, but we thought a team would represent more diverse experiences with pets, which could be communicated through discussion*
9	How many murders were	14,8	Source: FBI Crime

	officially registered in the US in 2013?	27	Statistics. *This question was added part-way through the experiment. This item was selected because we did not expect that there would be much unshared information to pool through discussion. Hence, this would serve as a no advantage for discussion "control question"*
10	In 1900, the percentage of the total US population that was aged 65 and over was 4.1%. What was that percentage in 2013?	13.1 %	Source: Administration on Aging. *This question was added part-way through the experiment. This item was selected because we did not expect that there would be much unshared information to pool through discussion. Hence, this would serve as a no advantage for discussion "control question"*

Question 7: Raven's Progressive Matrix (In order to solve the problem, participants needed to recognize that the lines are rotating, one small segment at a time, and choose the one of 8 images that completes the pattern if used to fill in the lower right block. First in the top left image, the lower right hand line segment rotates 90 degrees. This forms a new picture represented in the middle of the top row. Then the upper left hand line segment (i.e. the line segment directly opposite) in that new image rotates 90 degrees to form the image in the top right. Exactly the same rule applies to the middle row as we move left to right. As such, the correct answer is E.**)** *Our rationale for expecting discussion would enhance accuracy for this question is that the problem requires several logical inferences to reach correct answer, thus discussion could improve performance through error-correction.*



10. Empirical Study: Results. [Data analysis is still underway, so we only present some example results] Here we consider our initial analyses of human performance in estimating the height of the highest manmade structure (2722 ft, the Burj Khalifa in Dubai) and the number of station stops on the Chicago "L" Subway System (145 stops, in 2014 when the estimates were made). We chose these two items because the estimates exhibit different distributional properties, and pose an initial test of the sensitivity of our analytic calculations.

First, we consider the distributions of unaggregated individual estimates, focusing on the final Round 3 judgments, from each of the 3 judgment conditions (Solo, Judgment-Feedback, Discussion-Feedback). For Question 1 (height of tallest building), there is much more convergence in the Discussion-Feedback condition than the other two; for Question 4 (Number of L Stops), convergence is high in both feedback conditions, but not in the Solo condition. For Question 1, estimates are slightly more accurate for Discussion-Feedback, than for the other two conditions; for Question 4, there is no discernable difference in accuracy across the 3 Group Process conditions.

FIGURE #A: Tallest Building Individual Estimates: Histogram density functions, comparing 3 Group Conditions on Round 3 (Solo = Solo, Basic = Judgment-Feedback, Enhanced = Discussion-Feedback)

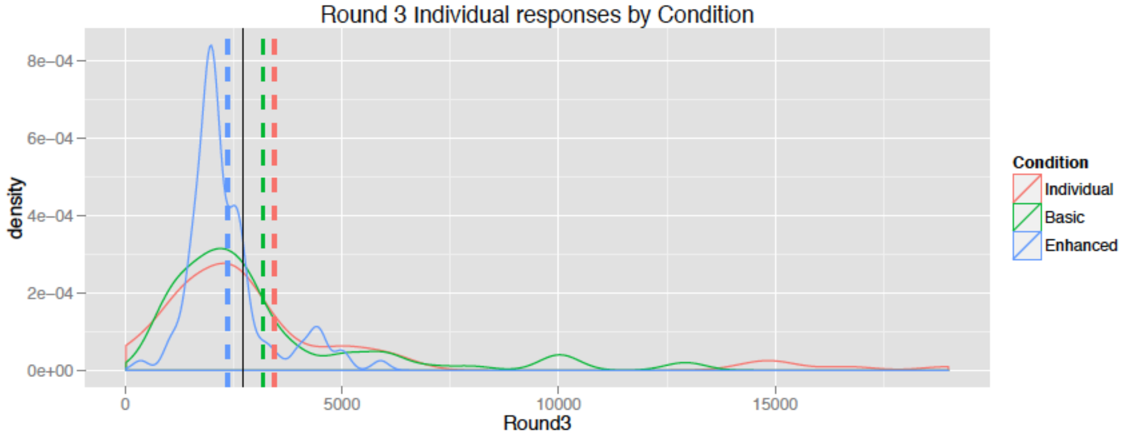
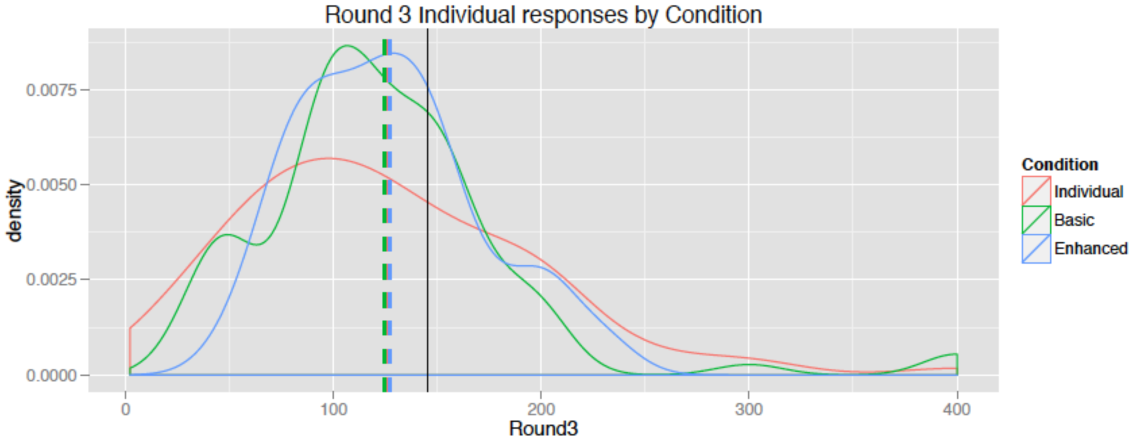
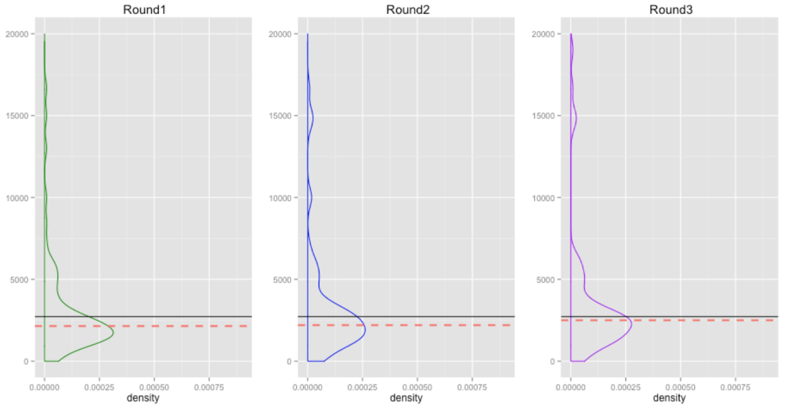


FIGURE #B: Number of L Stops Individual Estimates: Histogram density functions, comparing 3 Group Conditions on Round 3 (Solo = Solo, Basic = Judgment-Feedback, Enhanced = Discussion-Feedback)

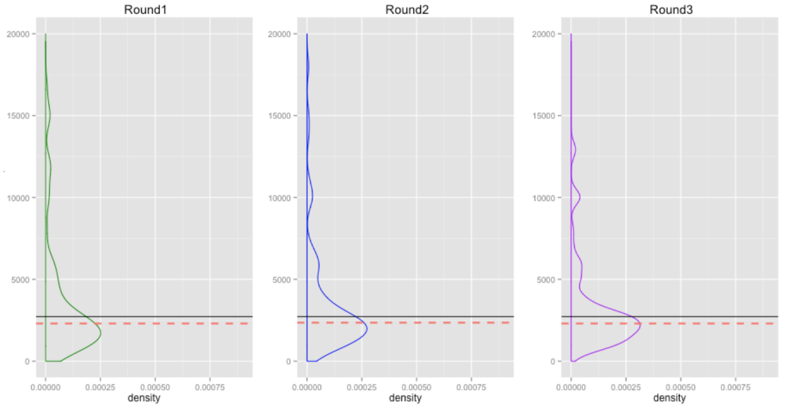


Second, we consider the distributions of individual estimates across the 3 rounds of judgments, separated by Judgment Condition. One obvious pattern is the convergence of estimates in the two Feedback conditions, especially strong in the Discussion-Feedback condition; while there is little convergence in the Solo condition. A more precise measure of within team convergence is provided by a table of the average standard deviations (across the 3 members' estimates within each team). All of these measures and displays support the conclusion that feedback on other team members' estimates or estimates + reasons produces more convergence than in the independence individual judgment, Solo condition. **We interpret all of these results as evidence for large discussion and persuasion effects in these factual belief judgment tasks.**

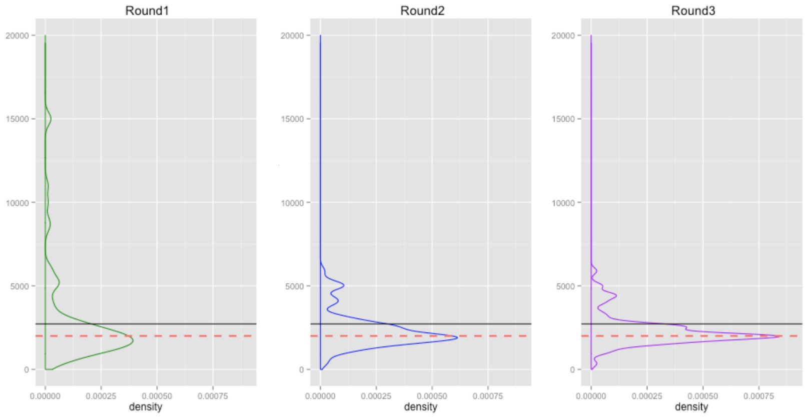
FIGURE #: Tallest Building Individual Estimates: Histogram densities, across 3 judgment rounds, for each Group Process condition.



Solo Condition Individual Responses < 20,000 across Rounds

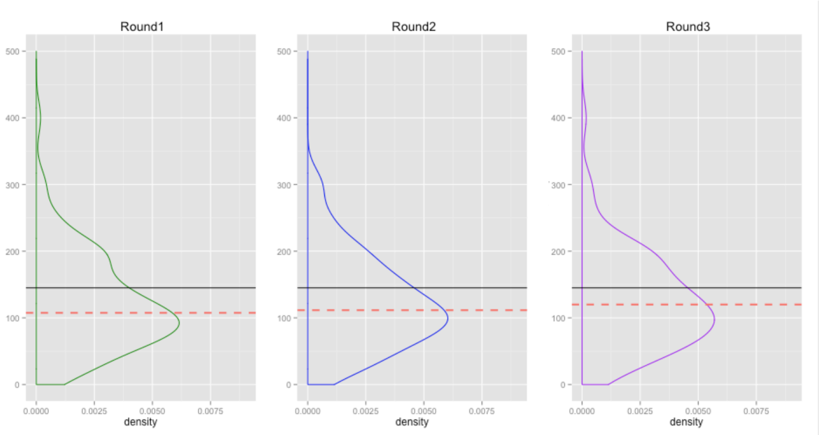


Basic Condition Individual Responses < 20,000 across Rounds

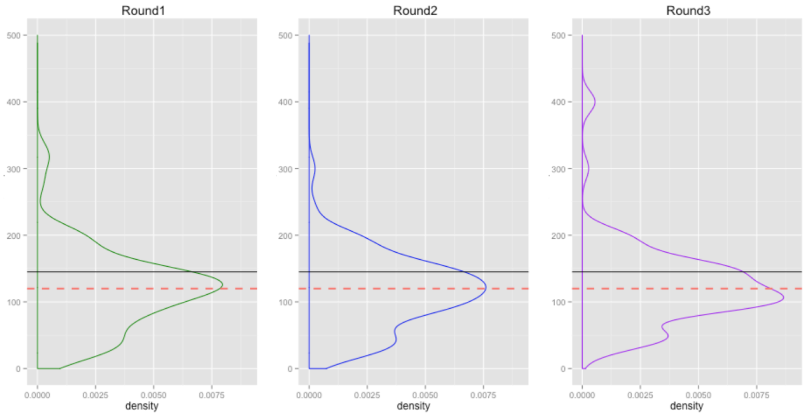


Enhanced Condition Individual Responses < 20,000 across Rounds

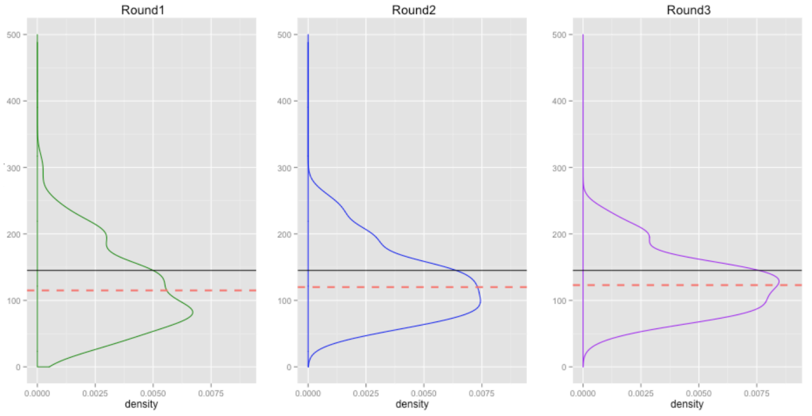
FIGURE #: Number of L Stops Individual Estimates: Histogram densities, across 3 judgment rounds, for each Group Process condition.



Solo Condition Individual Responses < 500 across Rounds



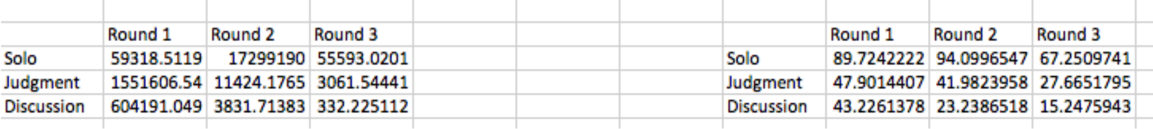
Basic Condition Individual Responses < 500 across Rounds



Enhanced Condition Individual Responses < 500 across Rounds

TABLE #. Measures of Convergence: Average Within-Team Standard Deviations

Q1: Tallest Building **Q2: Number of Subway Stops**



	Round 1	Round 2	Round 3
Solo	59318.5119	17299190	55593.0201
Judgment	1551606.54	11424.1765	3061.54441
Discussion	604191.049	3831.71383	332.225112

	Round 1	Round 2	Round 3
Solo	89.7242222	94.0996547	67.2509741
Judgment	47.9014407	41.9823958	27.6651795
Discussion	43.2261378	23.2386518	15.2475943

Third, we address the question of accuracy in group judgments with the calculation of a "Mean Absolute Proportion Error" score (MAPE = lestimate – truthl/truth; Gurcay, et al., 2014; Hoover, 2009) and use that measure to assess accuracy in the 3 Judgment conditions. Note there is a bias to over-weight large values in any analysis, and that bias is not corrected by using the standardized MAPE accuracy score. This occurs because for most of the quantitative estimates, there is a minimum value of zero, but the top of the scale is unbounded (so, for example we observed estimates of the height of the tallest building over 80,000 ft and of the number Chicago L stops over 500). The standardized accuracy score has a bound of 1.0, for underestimates (e.g., for an estimate of 0, if the truth is 2722 ft, MAPE score = (l0 – 2722l/2722) = 1.0), but no bound for over-estimates (e.g., for an estimate of 80,000, MAPE score = (l80,000 – 2722l/2722) = 28.4).

FIGURE #A: Height of Tallest Building Individual MAPE (accuracy scores): Density histogram

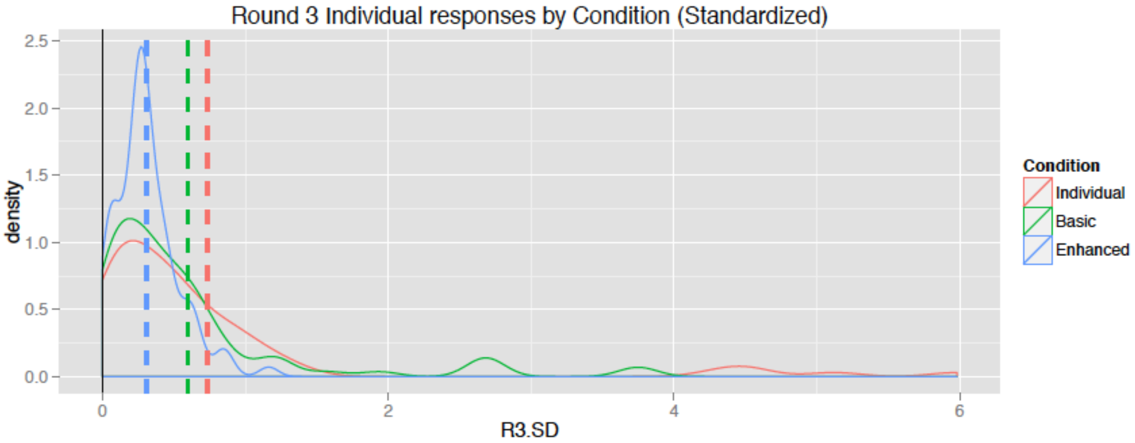
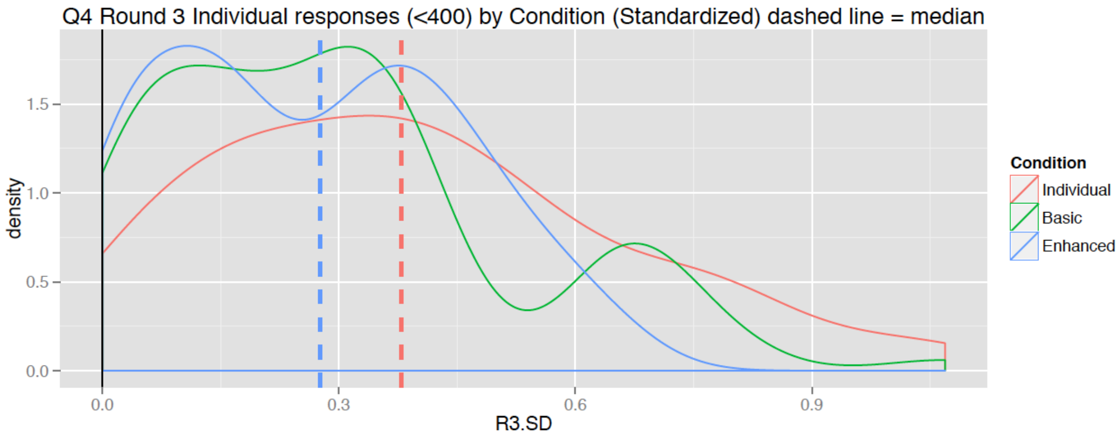
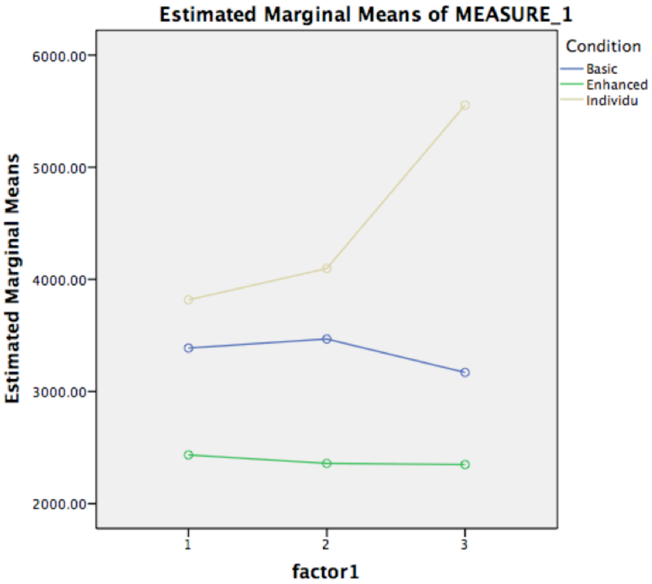


FIGURE #B: Number of L Stops Individual MAPE (accuracy scores): Density histogram



An analysis of variance was conducted to test for differences in average accuracy (MAPE standardized error scores) across judgment Rounds and across the 3 Group Process conditions. For the Tallest Building Question, the marginal means for the 9 conditions (3 Rounds X 3 Group Process conditions) are displayed below

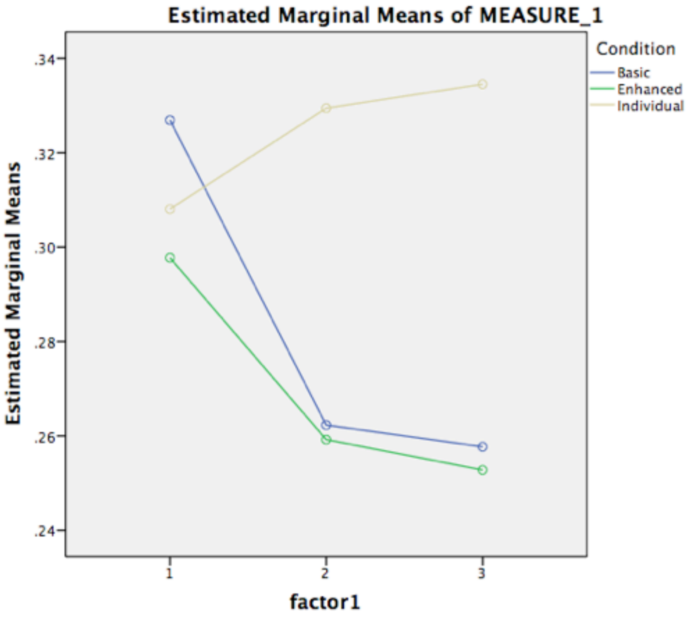
FIGURE #A: Tallest Building Question Marginal Mean Error Scores, for 3 Rounds ('factor1') X 3 Group Process Conditions



The Discussion-Feedback condition has the lowest error scores, the Judgment-Feedback condition is similar, but with higher error rates. The Solo condition is least accurate, with the highest error scores, increasing across the 3 Rounds (not predicted). Thus, the marginal mean MAPE scores are consistent with our hypothesis that the Discussion-Feedback and Judgment-Feedback conditions produce more accurate judgments than the Solo condition (the F-test statistic for the interaction between Rounds X Group conditions was F(2, 85) = 1.90, p < .05; the global F-test for differences between the Group conditions was F(2, 85) = 1.72, n.s.).

For the Number of Subway Stops Question, the plot of marginal means is also consistent with our hypothesis that feedback (Judgment-Feedback and Discussion-Feedback) group processes would be more accurate than the Solo process. Both feedback conditions showed decreases in average error rates across the 3 judgment rounds, while the Solo condition showed no improvement across rounds.

FIGURE #B: Number of Subway Stops Question Marginal Mean Error Scores, for 3 Rounds ('factor1') X 3 Group Process Conditions



An analysis of variance, testing the reliability of the suggestive interaction pattern was <u>not</u> significant (F(2, 86) = 1.009, n.s.), nor was the direct test of differences in average error rates across the 3 Group Process conditions (F(2,86) = 0.60, n.s.).

Our conclusions concerning accuracy of group estimates are disappointing. We find definite Wisdom of Crowds effects, such that the average error score (MAPE) for groups is less than the average error for individuals across all our questions. However, we did not observe consistent differences across 10 questions, in the relative performance of the 3 Group Process conditions. For a small majority (7 out of 10) the results were suggestive that the feedback Group Process conditions produced more accurate performances than the Solo condition (independent individual judgments); and for only 3 questions were statistical tests of the reliability of Group Process differences significant.

11. <u>What have we learned about Judgment-Feedback and Discussion-Feedback effects on judgments?</u>

The most dramatic effects of Judgment-Feedback and Discussion were on convergence of judgments towards consensus within teams. The pattern of reduced variance (within teams) is indubitable and shows up on every measure of convergence to consensus. It is clear if we look at the effects of Judgment-Feedback and Discussion-Feedback on extreme high and low judgments within each team; and it is irresistible in measures of within-team judgment dispersion (such as judgment standard deviations). Furthermore, it is clear that convergence within teams is greater with Discussion-Feedback than with Judgment-Feedback alone.

One current analysis task is to fit simple algebraic equations to individual 3-member teams to model the consensus process. Currently, weighted averaging models work fairly well (for the Judgment-Feedback and Discussion-Feedback process conditions). These models describe individual judgment shifts towards team-members' judgments, with impact of each member on the target individual as weighted by the distance between the two individual judgments. The further apart two judgments are, the less the influence of each on the other. We might call this model a center-of-gravity calculation, by analogy with the influence of physical objects' gravitational fields on one another's movements.

The effects of Judgment-Feedback and Discussion-Feedback are <u>not</u> clear on accuracy measures. Our conclusion is that accuracy increased for 6 out of 10 questions, given either Judgment-Feedback or Discussion-Feedback. This was true on both questions that *ex ante* we expected to show Discussion-Feedback effects and those we did not expect to show Discussion effects. In short, we failed to answer the question what types of questions will show a positive, advantage in accuracy following information-pooling, either of estimates (Judgment-Feedback) or of estimates and reasons for the estimates (Discussion-Feedback).

12. <u>What have we learned about the nature of effective discussion?</u> We collected team members' communications to one another in the

Discussion-Feedback condition. (Analysis is underway of the contents, seeking factors that discriminate between more and less accurate performing teams.) We provide a few examples of the contents in the following tables.

Table #. Examples of Discussion Content from each of the 10 Questions

#	**Example Insights from each Question**
1	“I believe the tallest building is probably about 150 stories, and with the average height of a floor being probably about 9-10 feet, this brings the height to 1350-1500 feet. Then, many tall buildings have tall poles or extra height added to the top, probably amounting to another 15 stories of height, meaning an extra 140-150 feet roughly. Added, this gives about 1500-1650 feet.”
2	“The US can be roughly approximated by a rectangle whose width is 1500 miles from the Gulf to Canada, and whose length is 3000 miles (the length from LA to NYC). Area of a rectangle is a=l*w.”
3	“I know there are 4 Starbucks in Hyde Park alone. If this were consistent throughout every neighborhood of Chicago, that would be at least 120 Starbucks in this city. Taking into account at least 10-20 other major cities with similar populations plus all the other smaller cities and Starbucks along highways, I estimated 3000.”
4	“I realize I wasn't thinking about just how many L lines there are. I still don't think the number is quite as high as the 160 or so posited by my teammates.”
5	“Miami is very close to the coast, and likely receive Cuban, Mexican, and other immgrants [sic] from South America. New York has the largest population, so it's between the two.”
6	“I know going from Chicago to New York takes about 24 hours, but Chicago to Nashville is only about 8 hours. Given the distance between Wyoming and Florida, and all of the stops, I would say that it would take at least 2 days, which is 48 hours, plus stops in cities out of the way.”
7	“Going from left to right, the first two rows suggest that you rotate the bottom right and then the upper left. Going from top to bottom, the first two columns suggest you rotate the bottom left and then the upper right; applying these transformations on the last column and last row both give E, so that is most likely the pattern.”
8	“I've heard the gun deaths in America are about 30,000 per year ... it's mentioned after nearly every mass shooting and that half are suicides, half homicides.”
9	“There’s about 300 to 400 murders in big Metropolitan areas like Chicago and New York per year, then add in Detroit and that might be another 500 or so and then a bunch in LA, then a bunch in the rest of the country.”
10	“Our population has nearly tripled in size, so I multiplied 4.1 by 3 to get 12.3.”

Table #. Dialog from One Effective Discussion-Feedback Team Answering the Number of Subways Question

Round	Participant	Insight
1	1	"I think there are 6 lines: Pink, Red, Brown, Orange, Green, Blue. Every line probably has about 20-30 stops, so on average... 25 stops x 6 lines = 150 stations"
1	2	"7 lines, each line has about 15 stops"
1	3	"The CTA lines are Red, Green, Blue, Pink, Orange, Brown, Purple, and Yellow. The Yellow line only goes to Skokie (1 stop), the Purple goes up Evanston from Howard (probably around 6 stops), I'm estimating around 20 stops for the rest of the lines and adding 4 because the Green line has a lot of stops very close to each other on the South side."
2	1	"A lot of the stations are shared, especially in the Loop, so it's reasonable the number is smaller."
2	2	"Another member more specifically provided information in her answer, so I revised mine considering her info"
2	3	"Red, Green, Pink, Brown, Orange, Blue, Purple (~6), Yellow lines (1), estimating 20-30 stations each, but since many stations overlap, conservatively estimating 20 unique stations, give or take a few (bringing us to ~130)"
3	1	"Raised my answer a bit to reflect other people's."
3	2	"Same as last time and another participant"
3	3	"Red, Green, Orange, Pink, Blue, Brown have around 20-30 stops each, Purple around 6, Yellow has one." Some overlap, so estimating the Chicago-only lines to be 20 each, plus a few."

13. Implications for Political Persuasion. Before we begin any consideration of implications for a realistic, representative political process, we must emphasize that our artificial and highly-controlled empirical study is not similar to any current political decision process. We focus on 3-member groups answering narrowly-focused factual questions, where every member's objectives are aligned, cooperatively, by a common incentive payoff scheme (okay, not exactly, as the incentives were paid to individual, not group winners). Any representative political process (except perhaps some legal examples) involves more individuals, individuals with more diverse backgrounds, and less cooperative objectives, a much more extended process, and questions involving a mixture of preferences and beliefs.

Galton (1907) commented that groups making judgments of true states of the world were analogous to groups making political judgments: "The average competitor [in an accuracy contest] was probably as well-fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues." It is also true that many political opinions are based on a mixture of factual reasoning and preferential reasoning. For example, citizen's concerned about the gun control policies, think about what they prefer (e.g., "I like to own a gun and use it for both recreational and safety purposes") and factual matters (e.g., "Gun control regulations will restrict my use of guns for recreation; gun control regulations will increase violent crime rates and I will, personally, be more endangered without strict regulation"). Some political discussions may focus almost completely on factual issues (e.g., "Will a proposed gun ownership regulation actually reduce the rate of violent crimes?"). All of these comments are suggestive, but not conclusive that research on groups making belief judgments has some relevance to the general topic of political persuasion.

14. Compare Contrast Persuasion on Beliefs versus Preferences

1. Consensus process converges on a weighted average central tendency in both … outliers are weighted less than central members (e.g., a geometric mean calculation, the Social Averaging Algebraic Model)

2. Confidence is related stubbornness in both …

3. Polarization is dramatic in preferences (e.g., Kahneman, Sunstein, & Schkade), not so much in beliefs

4. "Pure Belief": facts are premises, inferences are logical

5. "Pure Preference": values are premises, facts are 'inputs' to inferences; inferences are based on consequential (probability[fact] * payoff[value]) or "deontological" rule-based reasoning

References (under construction)