THE UNIVERSITY OF CHICAGO

THE VALUE OF AN ENHANCED
INFORMATION POOLING PROCESS
IN GROUP ESTIMATION AND FORECASTING

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

JOHN GRAHAM BURROWS

CHICAGO, ILLINOIS

AUGUST 2014

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Table of Contents

LIST OF FIGURES iv

LIST OF TABLES vi

ABSTRACT viii

INTRODUCTION 1

TWO-STAGE MODEL OF COLLECTIVE JUDGMENT 3

TASKS AND INFORMATION 4

OPTIMIZING COLLECTIVE PERFORMANCE THROUGH SOLUTION INTEGRATION ... 8

OPTIMIZING COLLECTIVE PERFORMANCE THROUGH INFORMATION POOLING ... 13

Known Obstacles Or Threats To Effective Information Pooling 14

Improving The Sharing (Pooling) Process Via The Delphi Method 18

Enhancing The Delphi Method With Improved Information Pooling 20

THEORY DEVELOPMENT 21

STUDY 1: ENHANCING INFORMATION POOLING IN GROUP EVALUATIONS OF START-UP PROPOSALS 24

Study 1: Participants 26

Study 1: Materials 26

Study 1: Procedure 27

Study 1: Hypothesis 29

Study 1: Analysis And Results 30

Study 1: Discussion Of The Results 33

STUDY 2: CONTROLLED EXPERIMENTAL TEST OF DELPHI WITH ENHANCED INFORMATION POOLING 35

Study 2: Participants 36

Study 2: Materials 38

Study 2: Procedure 44

Study 2: Measures Of Accuracy 49

Study 2: Hypothesis 52

Study 2: Analysis And Results 53

Study 2: Discussion Of Results 79

CONCLUSION 83

The Recipe For The Information Pooling Advantage... 83

What Next? 85

REFERENCES 88

List of Figures

Figure 1. Box Plot Summary Of The Proportion Of Groups Members (In Each Team) Who Picked The Winner Of The Nvc Competition, Across Rounds 1, 2, And 3 Of Judgments, Comparing Delphi With Reasons Versus The Delphi No Reasons Teams. 31

Figure 2. Line Graph Summary Of The Average Proportions Of Group Members (In Each Team) Who Picked The Winner Of The Nvc Competition, Across Rounds 1, 2, And 3 Of Judgments, Comparing The Delphi With Reasons (Solid Line) Versus Delphi No Reasons (Broken Line) Teams. 32

Figure 3. Box Plot Summary Of The Confidence Adjusted Accuracy Scores Of Group Members (In Each Team) Who Picked The Winner Of The Nvc Competition, Across Rounds 1, 2, And 3 Of Judgments, Comparing The Delphi With Reasons Versus The Delphi No Reasons Teams. 32

Figure 4. Line Graph Summary Of The Confidence Adjusted Accuracy Scores Of Group Members (In Each Team) Who Picked The Winner Of The Nvc Competition, Across Round 1, 2, And 3 Of Judgments, Comparing The Delphi With Reasons (Solid Line) Versus The Delphi No Reasons (Broken Line) Teams. 33

Figure 5. Raven's Progressive Matrix Used As Question 7 42

Figure 6. Trimmed Box Plots Of Percent Absolute Error Data For Delphi And Individual Group Methods; By Round Of Judgments (Excluding, Approximately, The Top And Bottom 2% Of The Outlying Observations). 56

Figure 7. Trimmed Box Plot Summary Of The Values For Geometric Mean Of The 3-Member Groups' Percent Absolute Error Scores For Q1 (With Outliers, "Outside The Whiskers" Excluded). 60

Figure 8. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q2 With Outliers Excluded. 61

Figure 9. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q3 With Outliers Excluded. 62

Figure 10. Trimmed Box Plot Summary Of The Values For Geometric Mean Accuracy For Q3 With Outliers Excluded. 63

Figure 11. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q4 With Outliers Excluded. 64

Figure 12. Trimmed Box Plot Summary Of The Values For Geometric Mean Accuracy For Q4 With Outliers Excluded. 65

Figure 13. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q6 With Outliers Excluded. 67

Figure 14. Trimmed Box Plot Summary Of The Values For Geometric Mean Accuracy For Q6 With Outliers Excluded. 68

Figure 15. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q8 With Outliers Excluded. 70

Figure 16. Trimmed Box Plot Summary Of The Values For Geometric Mean Accuracy For Q8 With Outliers Excluded. 71

Figure 17. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q9 With Outliers Excluded. 73

Figure 18. Trimmed Box Plot Summary Of The Values For Geometric Mean Accuracy For Q9 With Outliers Excluded. 73

Figure 19. Trimmed Box Plot Summary Of The Values For Median Accuracy For Q10 With Outliers Excluded. 75

Figure 20. Trimmed Box Plot Summary Of The Values For Geometric Mean Accuracy For Q10 With Outliers Excluded. 75

List of Tables

Table 1. 8 Types Of Task According To Mcgrath (1984) 5

Table 2. Quadrant Ii: Choose Tasks From Mcgrath (1984) 6

Table 3. Demographics Of Study 2 Participants By Location 37

Table 4. Study 2 Questions 1 Through 10 (Excluding Raven's Question 7) 43

Table 5. Measures Of Accuracy (Based On Dummy Data) 51

Table 6. Outlier Estimates For Question 1 (Crl). 57

Table 7. Outlier Estimates For Question 1 (Drl). 57

Table 8. Median Round 3 Accuracy For Question 1 (Drl). 58

Table 9. Kruskal-Wallis Analysis Of Ordinal Ranks Only For Round 3 Estimates Of Question 1 (Crl And Drl). 59

Table 10. Ranges Of Numerical Answers To Q2. 61

Table 11. Kruskal-Wallis One-Way Analysis Of Ordinal Ranks Only For Round 3 Estimates Of Question 3 (Crl And Drl). 63

Table 12. Kruskal-Wallis One-Way Analysis Of Ordinal Ranks Only For Round 3 Estimates Of Question 4 (Crl And Drl). 66

Table 13. Kruskal-Wallis One-Way Analysis Of Ordinal Ranks Only For Round 3 Estimates Of Question 6 (Crl And Drl). 68

Table 14. Kruskal-Wallis One-Way Analysis Of Ordinal Ranks Only For Round 3 Estimates Of Question 8 (Crl And Drl). 71

Table 15. Ranges Of Numerical Answers To Q9. 72

Table 16. Kruskal-Wallis Ordinal Ranks Test Only For Round 3 Estimates Of Question 9 (Crl And Drl). 74

Table 17. Kruskal-Wallis One-Way Analysis Of Ordinal Ranks Only For Round 3 Estimates Of Question 10 (Crl And Drl). 76

Table 18. Examples Of Shared Insights 78

Table 19. Dialogue Among Members Of One Enhanced Delphi Team About Question 4 79

Table 20. Summary Table Of Results. 80

Abstract

There is a great deal of interest in how to leverage dispersed information and to fully utilize the expertise of individuals from different domains when they are brought together as a group. The Delphi method is one technique designed to facilitate information pooling and the utilization of collective intelligence. In either its classical or one of its many modified forms, the Delphi method has been with us for over 60 years now; yet it still remains unclear whether the technique does, in fact, serve to better exploit collective intelligence. We test an enhanced information-pooling version of the Delphi method that facilitates pooling of unshared, valid, judgment-relevant information. We compare three judgment conditions: (i) Individuals acting in isolation, repeating their estimates three times, serving as the experimental control condition; (ii) Basic Delphi—individuals estimate three times, and after the first and second round they receive a list of the other group members' estimates; and (iii) Enhanced Delphi—individuals estimate three times, and after the first and second round they receive a list of the other group members' estimates plus useful information shared by each other group member. We seek to answer the following question: Are estimates from the Enhanced Delphi Method more accurate than from Basic Delphi and from isolated individuals? For the purposes of assessing accuracy, we conduct statistical analyses comparing individuals, groups, and group averages.

Keywords: Information pooling, collective intelligence, Delphi Method, group judgment, group decision making, Wisdom of the Crowds

Introduction

Virtually every decision in business (and plenty outside of business) involves bringing together several experts to work together to solve a shared business challenge. These experts bring to bear on a problem one or both of local domain knowledge (e.g., knowledge of European versus Asian geography, culture, and population centers) and/or distinct functional expertise (e.g., expertise in accounting versus marketing).

Firms look to facilitate information pooling. For example, one firm's CEO might want the firm's management team to get together with the firm's investment bankers and/or the firm's strategy consultants to determine which company to acquire or which existing division to divest. Another firm might bring various constituents together to determine whether to launch a new product or instead to license similar technology from another firm. A Board of Directors aided by independent auditors might try to determine whether a firm's management team is attempting to hide evidence of malfeasance. Property developers might want to collaborate to determine where to situate a shopping mall. A team of VCs with different industry backgrounds might want to determine whether to invest in start-up company A from one industry or start-up company B from another. Apple might want to utilize the collective wisdom of its management team to predict future actions of a competitor such as Samsung or Google, or gauge the likelihood of legal litigation relating to patent infringement, for example.

In each of these cases, a group of individuals is brought together and asked to work towards a common end. It has become a cliché that we need cross-functional teams and diversity (whether ethnicity, nationality, gender, or merely geographic and/or industry focus) in the workforce because such diversity leads to better business outcomes. The often unstated and arguably naïve assumption on which such ideas is based is that “Two heads are better than one.”[1] While this purported nugget of wisdom has enjoyed sufficient support over the years to become a proverb, in practice it is not always true.

The proposed research explores methods to increase the efficacy and accuracy of predictions and solutions generated by a small team of problem solvers. While group versus individual performance has been studied extensively, we are still relatively uninformed about how to acquire and integrate dispersed information and fully utilize the expertise of individuals from different domains working together as a group. This is the problem of information pooling and the focus of this dissertation.

One collection of methods to improve group estimates and forecasts that has been studied extensively involves integrating individual solutions together “algorithmically” without direct interaction among the team members. Various individual solution weighting methods and voting schemes have been tested and substantial improvements in the collective solution have been achieved (e.g., Clemen & Winkler, 1999; Dawid, DeGroot, & Mortera, 1995; Graefe & Armstrong, 2011; Hastie & Kameda, 2005; Hogarth, 1977; Jacobs, 1995).

While there is published research articles that evaluates how to optimize the weighting of various inputs to a decision, there are a great many more unpublished articles and online discussions that speak to these issues. Perhaps the best source of active, timely, and intelligent

[1] This proverb is first recorded in John Heywood's *A dialogue conteinyng the nomber in effect of all the prouerbes in the Englishe tongue*, 1546: “Some heades haue taken two headis better then one: But ten heads without wit, I wene as good none.” A similar idea is expressed in *Ecclesiastes, 4:9*, in *Coverdale's Bible*, 1535: “Therfore two are better then one, for they maye well enioye the profit of their laboure.”

discussion of these matters is IARPA ACE's own website: http://www.iarpa.gov/index.php/research-programs/ace.[2]

In contrast to most ongoing research in this field, which focuses on the aggregation of individual solutions, we are interested in enhancing the exchange of information among team members to increase accuracy of a simple composite score such as an average or median accuracy. Although information pooling in small face-to-face groups has frequently been studied, there is little research on how to enhance information pooling to increase performance. This is the focus of the present research.

Two-Stage Model of Collective Judgment

To clarify our focus, we note that any collective judgment task can be decomposed into two stages that we will refer to here as Information Acquisition versus Solution Integration respectively. In the Information Acquisition Stage sharing (pooling) of information, evidence, and individual judgments occurs. In Solution Integration Stage individual team members' information, evidence, and individual judgments are combined either through a social or a 'mechanical' process. The outcome of the two stages is a unitary answer representing the collective judgment of the group. The Solution integration Stage has been studied extensively. Much effort has been expended trying to determine what aggregation rules maximize accuracy, or what voting schemes, statistical combination equations, etc., are optimal (Dawid, DeGroot, & Mortera, 1995; Hastie & Kameda, 2005; Hogarth, 1977; Jacobs, 1995). Because these processes have received the lion's share of attention from researchers, we shall review what is known in a

[2] The Intelligence Advanced Research Projects Activity (IARPA) invests in high-risk, high-payoff research programs to tackle some of the most difficult challenges of the agencies and disciplines in the Intelligence Community (IC).

little more detail before switching gears and focusing on the information acquisition and sharing processes that are the focus of our research.

Tasks and Information

We believe that effective information acquisition and pooling will produce a new generation of collective intelligence methods that will significantly increase performance over the previous developments in, mostly, information aggregation mechanisms. But, we also believe that improvements in acquisition and pooling will only produce substantial improvements in performance of certain tasks. Thus, we need to say a few words about the types of collective intelligence tasks that will show the greatest improvements.

Researchers who study group processes are unanimous in rejecting the idea that task performance is unaffected by characteristics of the task being performed, yet almost no serious effort has been undertaken to explore how task differences affect group task performance. McGrath (1984) is right to assert that the choice of task type in a given study "is often a matter of convenience and fairly arbitrary." McGrath urges us to take up the task of analyzing and classifying tasks in ways that relate meaningfully to how groups perform them. McGrath is not alone in recognizing this shortcoming in the literature and nor is he the only one to make a serious attempt at rectifying the shortfall; but his is among the most sophisticated attempts to formulate a coherent task classification system.

In his book *Groups: Interaction and Performance (1984)*, McGrath applauds the efforts of other researchers in attempting to provide a useful means of understanding task differences and relations among tasks and, in turn, their impact on group performance in undertaking them. McGrath uses the works of Shaw (1973); Hackman (1968, 1976); Hackman, Jones, & McGrath

(1967); Hackman & Morris (1975, 1978); Steiner (1966, 1972); Laughlin (1980); Davis, Laughlin, & Komorita (1979) as his jumping off point. McGrath extracts the main ideas from the work of his intellectual forbearers and integrates them into his own proposed typology of tasks—what he calls his circumplex model of group task types.

McGrath proposes 4 quadrants of task categories under each of which is 2 types of task. The following (Table 1) highlights these 8 types of task. The table is drawn from McGrath (1984):

Table 1. 8 types of task according to McGrath (1984)

Quadrant I: Generate		
Type 1.	Planning Tasks	Key notion: Action-Oriented Plan
Type 2.	Creativity Tasks	Key notion: Creativity
Quadrant II: Choose		
Type 3.	Intellective Tasks	Key notion: Solving problems with a correct answer
Type 4.	Decision-Making Tasks	Key notion: Preferred answer
Quadrant III: Negotiate		
Type 5.	Cognitive Conflict Tasks	Key notion: Resolving policy conflicts
Type 6.	Mixed-Motive Tasks	Key notion: Resolving pay-off conflicts
Quadrant IV: Execute		
Type 7.	Contests/Battles	Key notion: Winning
Type 8.	Performances	Key notion: Excelling

Quadrant II, Choice Tasks, is most relevant to our research given the nature of the questions used in Studies 1 and 2. Under quadrant II are two task types: intellective and decision-making. The former are tasks for which there is a demonstrably correct answer. The latter are tasks for which there is not a demonstrably correct answer, and which ask of participants that they work together to select, by some consensus, a preferred alternative.

McGrath suggests that intellective group tasks fall into one of three subsets: (i) those for which the correct answer is intuitively compelling once put forward (i.e., Eureka tasks); (ii) those for which there is a correct answer with a logical path to the solution, but it may be difficult to demonstrate the logic in a way that is intuitively compelling to all members of the group (we call these "demonstrable solution tasks"); and (iii) those for which the correct answer is based on a consensus of experts. It is important to note that this last subset of intellective task is little removed from tasks where the correctness of an answer is based on the consensus of the focal group itself, i.e., a jury. In such cases, the task is categorized as a decision-making rather than intellective task.

McGrath also proposes subsets for decision-making tasks but acknowledges they are "less clearly distinctive." The source of the "correct" answer is perhaps the best way to distinguish the three subsets of decision-making tasks: (iv) those tasks for which the "correct" answer based on shared cultural values; (v) those tasks for which the "correct" answer involves social comparison and other social influence processes; and (vi) those tasks for which the "correct" answer involves consensus attained by sharing relevant information.

Table 2. Quadrant II: Choose Tasks from McGrath (1984)

Quadrant II: Choose		
Type 3.	Intellective Tasks	i. Correct answer intuitively compelling ii. Correct answer but difficult to demonstrate iii. Correct answer based on a consensus of experts
Type 4.	Decision-Making Tasks	iv. Correct answer based on cultural values v. Correct answer involves social comparison vi. Correct answer involves consensus based on sharing relevant information

The task we asked participants to perform in study 1 was intellective and fell in the subset of (iii) because the "correct" answers in that study were based on a consensus of experts, i.e., the judges of the New Venture Challenge. The tasks in study 2 were also all intellective but this time they fell in the subset of (ii) because although all tasks had a demonstrably correct answer, in each case the demonstration was non-obvious.

We hypothesize that intellective choice and decision tasks will be most affected by changes in the information acquisition and pooling methods employed by groups performing these tasks. In addition, certain information distribution conditions must be met for variations in information-pooling to make a difference. We propose that there are three general categories of information, relevant to intellective problem-solving tasks: (i) data, evidence, reasons, or inputs to the problem-solving algorithm; (ii) the nature the specific solution algorithm or calculation; and (iii) unitary solutions generated by the integration of the first two categories. Solutions can take many forms from declarative reports to contracts to plans to achieve a goal or construct a product. But, in our research, for experimental simplicity, we will study problems with simple numerical or categorical answers (e.g., estimates of the height of the tallest building in the world; the major US metropolitan region with the non-native residents).

First, we take it as self-evident that the more types of relevant information there are, the greater the value that will accrue from improved information pooling methods. Note that some problems have very little shareable information. Consider the classic problem of estimating the number of beans in a jar. There is very little to share, aside from the individuals' estimates (solutions). But, in contrast, for many practical problems that arise in business, e.g., evaluating a business proposal, estimating the completion date for a multi-component project, valuing a complex entity like a company, there is a great deal of non-solution information to share.

Second, the other obvious dimension describing information concerns its distribution across group members. In some cases, all relevant information (data, inputs, and solution algorithm) might be known by all group members. However, a fully-shared distribution of information is almost unheard of (perhaps the only clear example, is the American trial jury, where great care is taken to provide all members of a jury with all the evidence and procedural instructions). The norm in most important endeavors is partial sharing. For example, most problem-solving teams in business settings are composed of members who share some basic information about constraints on the solution, with some common background concerning some of the inputs to the ultimate solution. In practice team members are selected precisely because individuals have different complementary task-relevant data, knowledge, and problem-solving skill sets.

If we combine these two dimensions: (i) variety of task relevant information; and (ii) distribution of the information across members, we can see that the conditions under which effective information sharing will matter the most are those where there is a broad variety of task-relevant information and where that partially-shared information is widely distributed across group members. Thus, the aspiration of the present research is to design intellective group problem-solving tasks that involve a variety of types of information and where it is likely that information is distributed across group members at the start of the group problem-solving process.

Optimizing Collective Performance Through Solution Integration

Two general methods or institutions have been used to try to increase collective performance (accuracy in the tasks being studied), above and beyond the performance of (i) unstructured face-to-face groups, by improving the solution integration process: (ii) nominal statistical groups and voting electorates; (iii) structured methods to integrate opinions (e.g., Delphi Method, Prediction Markets). Face-to-face discussion is self-explanatory, with group discussion and decision/aggregation rules typically being generated in an *ad hoc* fashion.

Nominal statistical groups and voting electorates involve minimal social interaction (e.g., the many examples in Surowieki's *Wisdom of Crowds*). This idea of a nominal group of independent judges or decision makers, was originally developed by Delbecq & VandeVen (1971). It involves individual members of a nominal group voting on a decision. The number of votes each solution receives is totaled, and the solution with the highest (i.e. most favored) total ranking is selected as the final decision.

Refinements to this basic idea include defining a group's performance based on averages, sometimes averages weighted by expertise or confidence. Hill (1982) reviewed group versus individual decision making research and suggested that there was an urgent need to be more rigorous and systematic in the manner with which comparisons are made between individuals and groups. She suggested such comparisons can be made in one of four ways, but that researchers often mix them up and/or neglect to specify which is their focus: (1) group versus individual; (2) group versus the most competent individual in an aggregate; (3) group versus pooled responses of an aggregate; and (4) group versus math models of performance.

Without such precise terminology, Hill notes that erroneous conclusions have been drawn about the relative performance of groups and individuals. Davis-Stober, Budescu, Dana, and Broomell (2014) provide a comprehensive review and conclude that statistical aggregates (unweighted means, or means weighted by prior accuracy or confidence) consistently outperform "randomly" sampled group members and often outperform even the best members.

Structured methods relying on social interaction to integrate opinions deserve a little more explanation. Such methods include the Delphi Method, Prediction Markets, as well as mixed approaches best represented by the Intelligence Advanced Research Projects Activity's[3] Aggregative Contingent Estimation (IARPA-ACE) competition and one of the high performing teams in that competition: the University of Pennsylvania's Good Judgment Project.

The Delphi technique has been around since the mid-1950s when it was invented at the RAND Corporation to make strategically-important estimates such as the number of atomic bombs needed to reduce conventional US munitions output to a quarter of its current output (Dalkey, Helmer, & CA, 1962). Delphi was introduced to a wider audience with the publication of Linstone and Turoff's edited collection of essays on the subject in the 1970s (Linstone & Turoff, 1975). Since then, the Delphi technique has become widely used for forecasting in a variety of disciplines and has itself evolved in a number of different directions.

In the basic version of Delphi—also referred to as Classic Delphi—participants enjoy full anonymity by utilizing pen names, knowing only that other participants are members of their own peer group. Other variations of the Delphi method include using computer-mediated groups. Still others relax the anonymity requirement by making interactions face-to-face; while some use a moderator/facilitator, sub-groups, and/or anoint experts within groups. In a Delphi group it is always the case, however, that individuals provide initial forecasts, receive feedback in the form of some kind of summation of the team members' responses, and then have the opportunity on a

[3] According to an article in Wired Magazine, IARPA is "[m]odeled after DARPA — and [is] often referred to as 'DARPA for spies' — the agency never had the public breakthroughs of its military successor; IARPA never put any robotic cheetahs or mind-controlled machines on display, the way DARPA did. But … IARPA did pursue far-reaching investigations into everything from crowdsourcing to quantum computing, all in the name of providing 'the U.S. with an overwhelming intelligence advantage over our future adversaries,' as the agency repeatedly put it." from http://www.wired.com/dangerroom/2012/04/lisa-porter-iarpa/

round-by-round basis to iteratively revise their previous forecasts with the expectation that group performance improves and converges from round-to-round.

In spite of Delphi being compared with other techniques in circumstances where it is unclear what, if any, benefit Delphi could bring to bear on a given forecasting question, Delphi has, at times, been shown more effective than other techniques (Rowe & Wright, 1999; Woudenberg, 1991). As we interpret these empirical results—mixed as they are—we are mindful of what Rowe & Wright (1999) wrote: "…[t]here are theoretical and empirical reasons to believe that a Delphi conducted according to 'ideal' specifications might perform *better* than the standard laboratory interpretations." This means that Delphi offers a far more efficacious approach to answering certain kinds of questions than is widely thought.

Graefe & Armstrong (2011) analyzed the accuracy of three structured group decision-making approaches (nominal groups, Delphi, and prediction markets) and compared them to traditional, unstructured face-to-face meetings. They stated no directional hypotheses on the relative accuracy of the three structured approaches, but they did expect the accuracy of the three structured approaches to be higher than for unstructured face-to-face interaction. They sent their study design to experts for validation. Participants were 227 students who were divided into 44 groups (11 for each of the 4 methods). Most groups consisted of 4-6 participants. The authors found no statistically significant differences between the four groups.

Another structured approach, relying on prediction markets, has ebbed and flowed over time. They were popular in the late-1800s and early-1900s and are enjoying a resurgence in popularity right now (Rhode & Strumpf, 2004). According to Chen & Plott (2002) an internal market at HP, with the purpose of predicting future product sales beat the company's official forecasts 6 out of 8 times. Green, Armstrong & Grafe (2007) comment that researchers are also doing more in this area signaled by the inauguration of the *Journal of Prediction Markets* in 2007. In addition, until possible "financial irregularities" resulted in closing, Intrade (www.intrade.com) was a viable public prediction market. Inklingmarkets (www.inklingmarkets.com) is an example of a company that provides technological support to enable other firms to deploy prediction markets internally and, though still privately held, seems to be doing quite well.

Simply put, participants in prediction markets buy and sell contracts, just as do traders in real markets, especially markets for options. Wolfers and Zitzewitz (2006) tabulated three different types of contract: binary option, index futures, and spread betting. Each is tailored to make a particular kind of prediction. In a binary option market, the price at which a contract most recently traded (or an average of the most recent prices) can be interpreted as the market's assessment of the probability that the event will occur. For example, if a contract will pay $1 in the event of the U.S. sending troops into Nigeria to confront Boko Haram before the end of 2014 and nothing otherwise, and the contract last traded at 22 cents, then the market's assessment is that the likelihood of that event is 0.22 or 22%. Prediction markets are not our focus here, but they are interesting in that they provide further support that a structured group's accuracy can be improved, to some degree or other, over traditional, *ad-hoc*, face-to-face interaction alone.

The Aggregative Contingent Estimation (ACE) project, sponsored by the Intelligence Advanced Research Projects Activity (IARPA) has the goal of "dramatically enhance[ing] the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts."[4] The ACE program involves empirical testing of forecasting accuracy

[4] http://www.iarpa.gov/solicitations_ace.html

against real events. Examples from ACE's website include: The UK will leave the European Union before the end of 2012; Germany's renewable energy (solar, biomass, wind and hydro power) will account for 22% of their total energy usage in 2012; and Groupon will file for bankruptcy by January 15th 2013.[5]

IARPA specifically asks teams to improve group performance in the following areas: (1) efficient elicitation of probabilistic judgments, including conditional probabilities for contingent events; (2) mathematical aggregation of judgments by many individuals, based on factors that may include past performance, expertise, cognitive style, meta-knowledge, and other attributes predictive of accuracy; and (3) effective representation of aggregated probabilistic forecasts and their distributions. Current teams are affiliated with a range of research and educational institutions including George Mason University, University of California, University of Maryland, Harvard, MIT, and the University of Pennsylvania. Different teams are taking a variety of approaches. The team from George Mason University is focused on countering biases like group-think, anchoring, and overconfidence via prediction exchanges, conditional forecasts, and Bayesian updating.[6] The University of Pennsylvania's ACE team has been extremely adept at translating their involvement in ACE into a well publicized project—called the Good Judgment Project—and has generated a solid pipeline of thoughtful and scholarly research articles including Gürçay et al. (2014) and (Mellers et al., 2014).

Optimizing Collective Performance Through Information Pooling

Although it seems intuitively plausible that improvements in information pooling would increase the accuracy of group, a highly visible experimental study recently concluded that

[5] http://www.forecastingace.com/aces/examples.php
[6] http://c4i.gmu.edu/pdfs/00_ace_mason_kickoff_v3.3.pdf

increasing information pooling can actually degrade performance on estimation and prediction tasks. Lorenz et al. (2011) imply that information pooling does not occur; or if it does occur, it degrades performance by increasing inter-dependence and redundancy between individual judgments. This provocative claim—provocative at least to those who believe that two heads are better than one—prompted a vigorous response from Mellers and her colleagues (Gürçay et al., 2014).

Where do we stand on this question? We believe that effective information pooling is possible. But we also believe that it improves judgments only in regards to certain problem types, and then, only when the team is composed of individuals who have something informative to share with one another. Notwithstanding the challenge, we build upon prior work to coax out the elusive improvements in performance from enhanced information-pooling in a Delphi structured group process.

Known Obstacles or Threats to Effective Information Pooling

There are well-known obstacles or threats to effective pooling that have been identified across several related research domains. These include: The Hidden Profiles bias (Stasser & Titus, 1985); the "Common Knowledge Effect" (Gigone & Hastie, 1997); "Group Think" (Janis, 1972); and "Associative Blocking" (Denton & Kruschke, 2006). We now provide a very brief overview of each of these:

Stasser and Titus (1985) introduced the paradigm that is sometimes referred to as the Hidden Profile Method. Larson (2010) does a particularly good job of summarizing the pertinent facts and takeaways of the method. He notes that a typical example that demonstrates the power of the hidden profile in a financial investment task (see page 178 in Larson's book for additional details). In such a Hidden Profiles task, participants, in perhaps four-person groups, play the role of investors considering a pair of startup companies, A and B, each of which is launching a new product. Packets of information contain decision-relevant information for participants to review. The manipulation is that information briefing packets differ between subjects. Some information is contained in all packets (fully shared information) and some is in only one packet (unique information), while most information is partially-shared by two or three members. Members study their individual packets and packets are collected.   Then group members discuss and decide on a course of action, an investment, by consensus.

Consider a specific example involving 18 pieces of information, 6 might imply Company A will be more successful, whereas 12 might imply Company B will be more successful. (Obviously, it is important that each piece of information is equally valid, important, and memorable.) As a consequence, the full set of 18 information items implies that B is the best investment. Representative samples of the information for group members would create what has been called a Manifest Profile (in which the better choice would be apparent to each individual prior to discussion). But, alternatively a Hidden Profile can be created by distributing information so that the correct investment is unlikely to be perceived from reading the individual packets alone, but can only come out following a full discussion. In other words, in a Hidden Profile set-ups an initial bias *against* the best option, by insuring the positive information about the inferior alternative is widely shared and the positive information about the superior alternative is initially unshared.

As already noted, Stasser and Titus (1985) were the first to examine the effect of hidden versus manifest profile information distributions on group decision-making. Their specific task was political, and involved three choice alternatives (three political office seekers), but was otherwise very similar to the investment dilemma described by Larson (2010). In their control condition, all packets contained all the available information. This created a manifest profile (without unshared information) and a rational group member would prefer the correct solution (candidate). In a second condition, they created a hidden profile in which all packets favored the same inferior candidate. Their third condition involved a mixed hidden profile, one in which two packets favored one inferior candidate and two favored a different inferior candidate (note there were three choice alternatives). Stasser and Titus (1985) found that 83% of groups in the manifest condition chose the best candidate, but only 18% of the groups in the two hidden profile conditions chose correctly. This was their finding in spite of the fact that groups in all conditions collectively held the same information about all candidates, i.e., all that differed was the way in which information was distributed across members prior to their group discussion.

Hidden profiles have been shown to impair decision making across a range of arenas, including about financial investments (Hollingshead, 1996; Kelly & Karau, 1999; McLeod, Baron, Murti, & Yoon, 1997), determining which suspect is most likely to have committed a murder (Stasser & Stewart, 1992; Stasser, Stewart, & Wittenbaum, 1995; Stewart & Stasser, 1998), and which job candidate is best (Scholten, van Knippenberg, Nijstad, & De Dreu, 2007; Schultz-Hardt, Brodbeck, Mojzisch, Kerschreiter, & Frey, 2006).

It should also be noted that these findings have been replicated outside of the lab. Christensen, Larson, Abbott, Ardolino, Franze, & Pfeiffer (2000) found trained professionals fall prey to hidden profiles. They looked at three-person teams of physicians diagnosing two complex patient cases (one involving Parkinson’s disease and the other lupus erythematosus). Each physician viewed a video recording of an interview with the patient. There were two conditions: manifest profile and hidden profile. Team diagnoses in the manifest condition were

100% accurate versus 71% accurate in hidden profile conditions (a statistically significant difference between conditions).

A closely related phenomenon is called the Common Knowledge Effect (Carley, 1986; Gigone and Hastie, 1993). The more members of a group with knowledge of a specific item before the group's discussion, the more discussion will be devoted to that item. Furthermore, the more discussion, the greater the impact of that specific item on the group's final decision. Thus, with the distributions of information in a Hidden Profile task, the misleading information, that is shared more broadly, will have a greater impact on the final decision, producing errors in the group choice.

The associative learning effect known as Associative Blocking occurs when a learned cue to predict an outcome is paired with a novel cue. Research with human and animal subjects has found that they tend not to associate the novel cue with the outcome; or in other words, learning about the cue has been blocked (Denton and Krushke, 2006). This finding is important because it contradicts many models of learning in which associative strength is incremented directly by the co-occurrence of cue and outcome (see Rescorla and Wagner, 1972, and Mackintosh, 1975, for the underlying theory).

Groupthink is said to occur within a group when a desire for harmony in the group results in over-conformity, herding, and an irrational decision making process producing sub-optimal outcomes. The idea is that group members try to minimize conflict and reach a consensus decision without giving much, if any, airtime to alternative viewpoints. The group accomplishes this by actively suppressing dissenting viewpoints. The myopic movement to reach consensus also produces overconfidence and an exaggerated impression of the group's decision making capacities. In extreme cases, groupthink may produce dehumanizing actions against an "outgroup" or a dissenting minority within the larger group (Janis, 1972).

Improving the Sharing (Pooling) Process via the Delphi Method

We resist the trend among researchers to focus on aggregating individual solutions, and instead shift our attention to the information pooling process. Our specific research question is whether an enhanced information pooling method increases accuracy? We address that question by testing alternative information-sharing processes. Through our research we hope to define methods to improve the information-sharing (pooling) process. We do so by testing an enhanced information-pooling method within the framework of the Delphi Method.

We believe the Delphi Method represents a robust and efficacious means of facilitating not only the integration of judgments (not our focus here) but also information pooling (our focus) and is both under-appreciated and under-utilized as a technique for increasing collective performance. We provided a brief introduction to the Delphi Method above, but will now more thoroughly explicate on the technique. The Delphi technique has been used in a wide range of applications. While the technique originated at the RAND Corporation at a time when RAND was focused almost exclusively on military issues, the Delphi technique is now mainly used in business and government policy applications (Landeta, 2006). According to Green, Armstrong, & Graefe (2007), business applications of Delphi have included forecasts relating to the "Argentine power sector, broadband connections, dry bulk shipping, leisure pursuits in Singapore, rubber processing, Irish specialty foods, and oil prices." Forecasts of technology are also popular (i.e., about intelligent vehicle-highway systems, industrial robots, intelligent internet, and technology in education). And another arena in which Delphi is used is in relation to forecasting social issues; specific uses have included predicting the “urban future” of Nanaimo in British Columbia, and the future of law enforcement.

The Delphi technique has been continually modified since its early use in the 1950s. The Delphi design adopted in a given application has more often than not been situational: “guided by the research problem rather than by the requirements of the method” (Felicity & Sinead, 2011). There are now as many as ten main versions of Delphi, including: classical, modified, decision, policy, real time, e-Delphi, technological, online, argument, and dis-aggregative.

Some of the categories of Delphi are specific techniques in their own right, while others merely append new technologies, including web pages, and real-time electronic communication to the original Delphi design. There is also no standardization in terms of the number of rounds, the degree of anonymity enjoyed by participants, and the kind of feedback given to participants between rounds whether qualitative or quantitative (Felicity & Sinead, 2011). As a consequence of the large number of variations of the Delphi technique, we are more than sympathetic to the opinion of (Rowe & Wright, 2011) who suggest it is better to speak of Delphi technique*s* in the plural.

While the Delphi technique in either its “classic” form or one of its modified forms has been *used* extensively, no form of Delphi has been *studied* rigorously. To be clear: there have been empirical evaluations claiming to assess the effectiveness of Delphi, but as noted by Rowe & Wright (1999), “Most of such studies have used versions of Delphi somewhat removed from the ‘classical’ archetype of the technique.” This has made it challenging to generalize the findings from one Delphi study to another, either in terms of what makes one form of Delphi more effective than another, or in terms of how Delphi compares with other group decision-making techniques.

Those technique-comparison studies that have been completed have often utilized simplistic questions, variable and even non-existent iterative feedback, unsystematic information sharing, and students or layperson subjects (none of which would have been considered appropriate by Delphi's original adherents). It would have been more constructive to focus on the *process* by which Delphi (in one form or another) might be advantaged (or disadvantaged) against another Delphi form, and/or other structured group-decision techniques. Furthermore, the designs of many of these studies mean that when Delphi (in one form or other) faces off against other structured group techniques it does so disadvantaged vis-à-vis the other technique(s) in some manner. Making things worse is the fact that the authors rarely recognize such disadvantage and it therefore often goes unacknowledged.

Enhancing the Delphi Method with Improved Information Pooling

Bolger & Wright (2011) provide some guidance in terms of a research agenda to assess the efficacy of the Delphi technique in terms of exploiting collective intelligence. In their theoretically rigorous article—an article informed by lessons learned from various sub-fields within social psychology—the authors highlight the processes by which Delphi facilitates opinion change in groups. The authors focus on those processes that might reasonably lead to improvement in Delphi's accuracy (with the proviso that said improvement is at answering questions appropriate for Delphi). They identify "residual and normative and informational pressures towards consensus that potentially reduce process gain that might otherwise be achieved. For instance, panelist confidence may act as a signal of status rather than be a valid cue to expertise" (Bolger & Wright, 2011). The authors argue that it is imperative that "good cues" are provided to Delphi participants to produce better outcomes in terms of accuracy. And that such improved accuracy results from participants further from the "truth" changing their opinion more than participants who are closer to the truth, i.e., the less knowledgeable participants must change their opinions more than the more knowledgeable.[7]

Bolger, Stranieri, Wright, & Yearwood (2011) assess the impact of Delphi participants exchanging "rationales" or "reasons" among themselves in support of their first round answers to a given question. Specifically, Bolger et al. (2011) asked participants to predict the outcome of an Australian Rules Football match and, in the treatment group, also share a reason for their prediction of the match outcome, e.g., based on past and current form of the teams, availability of key players, home ground advantage, etc.

Theory Development

Building on the framework that is the basic Delphi Method, we plan to study a refined method—induced via instructions and practice—that better facilitates the pooling of unshared, valid, judgment-relevant information. We intend to compare 3 judgment conditions: (i) Individuals acting alone (I) in addressing a judgment task, then repeating their estimates three times.

(ii) Basic Delphi Method (B) – individuals in the Delphi estimate three times, and on each round they receive a list of the other group members' estimates; and (iii) Enhanced Delphi Method (E) - individuals in the Delphi estimate three times, and on each round they receive a list of the other group members' estimates plus useful information shared by each other group member.

[7] The authors also note that opinion change might be modeled in a stochastic form where the probability of a participant changing his or her opinion increases with distance from the truth.

The key differentiator between groups (note: the "I" or individual condition is primarily a baseline "control" treatment) and, we hope, the main contribution of this research, is demonstrating the value in sharing useful, actionable information between Delphi members such that group performance is increased. Our basic hypothesis is that final estimates from the Enhanced Delphi Method (E) are more accurate than final estimates from the Basic Delphi Method (B) and/or from Individuals Acting Alone (I).

Previous research, including that cited above, has questioned the value of exchanging reasons or rationales between group members; however, Rowe, Wright, & McColl, (2005) plausibly suggest the reason exchange typically fails to produce an improvement in outcomes is because, in most cases, the rationales exchanged between participants were poor, being more of a teleological rather than causal nature.[8] We are sympathetic to this point of view.

Bolger & Wright (2011) suggest that for reasoned feedback to be most effective—and to give Delphi the greatest opportunity to shine in a laboratory-based assessment—the following conditions must be met: (1) High level of expertise among participants; (2) variety in initial expert opinions; (3) the task must permit elaborate causal reasoning; (4) high extrinsic

[8] In a related study by Wentholt, Fischer, Rowe, & Marvin, (2010) participants were judged as providing uninformative reasons for their contrary views, and thus were unpersuasive in nudging others towards their viewpoint. Similarly—and writing of "egocentric advice discounting" in Judge Advisor Systems (JAS)—Yaniv found that a judge typically shifted a "token" amount—around 20% or 30%—towards his or her advisor's initial estimate (Yaniv & Kleinberger, 2000). In a more recent paper, Yaniv suggests that such "advice discounting" occurs because judges have access to their internal justifications for arriving at a particular decision and are therefore able to assess the strength of the supporting evidence for that decision. On the other hand, judges don't have access to their advisors' reasoning and therefore find themselves with less evidence justifying the advisors' decisions. Yaniv speculates that less knowledgeable judges could presumably retrieve less supporting information for their own opinion and therefore discount less (Yaniv, 2004).

motivation to perform the task resulting from monetary incentives or public recognition; and (5) a task involving as few as possible judgments per round.

According to Bolger and Wright (2011), most laboratory-based studies of Delphi tend to use "impoverished feedback" compared with actual applications of Delphi. They redirect our attention back to the original use of the Delphi method at the RAND Corporation in the mid-1950s.

In one of the original uses of Delphi, a detailed data set was elicited from subjects and fed back to them along with their initial judgments. As noted earlier, participants were asked to determine the number of atomic bombs needed to reduce US munitions output to a quarter of its current output. That said, focusing on the basic problem alone, misses the subtlety and complexity of what was asked of the participants. For in addition to answering the basic question, participants were asked to address the following: (1) Vulnerability of potential targets and their ability to rebuild; (2) analytic steps and computations to be performed (e.g., identification of optimal targets and estimation of minimum number of bombs on target to be 50% confident of reducing capability of those targets by one fourth); (3) suggestions of areas requiring further thought or research; and (4) arguments and justifications for both judgments and other data tendered (Dalkey et al., 1962).

The complexity of such questions is a far cry from a Delphi being used to answer an almanac question like "What is the height of the Sears Tower?" or to estimate the percentage of the US population over the age of 65 or who are college graduates (Graefe & Armstrong, 2011). While Bolger et al., (2011) takes a major step in the right direction—relative to other recent attempts to move the field forward—it seems that the authors still have not gone far enough in the direction of addressing richer, more complex estimation problems. Bolger and his co-authors fall short in three key ways: (1) the forecasting problem used (predicting the outcome of an Australian sports game) was not of the kind for which Delphi was likely to shine; (2) participants were not sufficiently practiced in the use of rationales; and (3) in no real sense were participants subject matter experts, except perhaps in that the problem asked so little of participants that any Australian would be deemed an expert for the purposes of the study.

Inspired by Bolger and his co-authors we aim to bring a little more empirical rigor to the table. To our minds, an appropriate assessment of Delphi must involve: (1) a suitably complex problem; (2) reliable use of rationales (i.e., ensure the participants are trained to share substantive and useful information about their answers to the questions being asked); (3) problems for which the relevant information is likely to be distributed and not completely shared across members of the group; and (4) if feasible use sufficiently expert participants (relative to the problem being asked) and ideally draw upon participants from distinct academic and/or functional areas.

Study 1: Enhancing Information Pooling in Group Evaluations of Start-up Proposals

Around the time we were writing the procedure for our first study, our theoretical discussions were focused on determining the kinds of problems/challenges for which information pooling ought to increase performance. At the time, it seemed clear to us that in a study exploring information pooling with a judgment, we needed to ask participants to make judgments about more than one item and ideally we also needed to include different types of information. We thought a complex problem involving small groups making judgments about which of three firms to invest in would fit the bill, especially as we had access to rich media including videos of actual pitch presentations, executive summaries, PowerPoint slides, and business plans from teams competing to win the Booth School's New Venture Challenge (NVC).

The NVC is a competition held annually at the University of Chicago's Booth School of Business, sponsored by the Polsky Center for Entrepreneurship. It begins as a two-quarter class in which Booth MBA students prepare a business plan that they ultimately present to a panel playing the role of a venture capital firm. Real prize money and equity is involved. The MBA students who choose to participate demonstrate their entrepreneurial skills and either have an equity stake and/or a management role in the early stage company they are pitching. Each NVC team is typically composed of 4 or 5 individuals. The judges act out of sense of community (they are typically Booth alumni now in industry roles and/or Booth professors) and out of self-interest; many of them are active angel or venture capital investors.

For the purposes of this study, we chose to use 3 teams from the 10 who competed in the 2009 NVC: NINE Naturals (who were the joint 1st place team that year and ranked by the judges as 2 of 10); CardTag/Mercadi (who were joint 2nd place with 3 other teams and ranked by the judges as 5 of 10); and Renovatio Labs (who failed to place along with 3 other teams and were ranked by the judges as 7 of 10).

Participants were provided background information on the Delphi technique, and were schooled in the value of providing clear and compelling rationales or reasons to support their decision. Then, as members of a Delphi, they were asked to complete a series of forecasting tasks relating to their assessment of the three NVC competitors used in the study. We used two conditions: Numeric Delphi (DN) and Delphi with Reasons (DR).[9]

[9] It is worth noting that in our second study we changed our terminology slightly and chose now to refer to the analogous conditions as Basic (B) and Enhanced (E) Delphi. Furthermore, in our second study we chose to include a baseline individual condition.

Study 1: Participants

Participants were drawn from a University of Chicago Booth School of Business class on Business Policy taught by Professor Harry Davis. These 137 participants were randomly assigned to groups in one of two experimental conditions: Numeric Delphi (DN) and Delphi with Reasons (DR). There were a total of 20 groups, 10 in each condition. Groups contained 6, 7, or 8 members. Of the 137 participants: 73% were male; 29% majored in science, technology, engineering, and/or mathematics (STEM) in college; 48% majored in business in college; and 35% intended to pursue a career in financial services.

Study 1: Materials

The principal stimulus in study 1 was a 35-page briefing document that contained PowerPoint slides and executive summaries from the 3 companies participants were asked to evaluate, and 3 10-minute video clips one for each of the companies presenting their value proposition in the 2009 New Venture Challenge at Chicago Booth.

All data collection was done with pen-and-paper and only later inputted to computer for analysis. Participants were asked to complete 3 rounds of evaluations of the 3 companies. Each time they were asked to rate the company proposal on a 5-point scale from 1 = very bad to 5 = very good. We also asked that they provide a 5-point confidence rating of their summary score.

The degree of information shared between group members from round-to-round differed according to experimental condition, but always included a histogram of the distribution of ratings within each group. Rationales or reasons, when they were shared, were photocopied from the original sheets completed by participants and were shared only in the Delphi with Reasons condition.

Study 1: Procedure

Delphi Background and Rationale Training for All Participants

All participants were provided background information about the Delphi technique, including its historical and recent applications, how it compares to other group decision techniques, etc. Furthermore, all study participants were told what would be asked of them as members of a Delphi. All participants were instructed about making first round forecasts, sharing those forecasts with their group-mates, and then iteratively updating and improving their forecast accuracy in subsequent rounds based on the numerical summation of their group-mates' forecasts as well as the supporting reasons provided by their group-mates. All participants learned the importance of providing compelling reasons for why they believed a certain outcome would transpire, and saw concrete examples of such. It was noted that causal rather than teleological reasons were preferred and participants were told that such reasons have been shown to best improve the outcome of a Delphi in subsequent rounds. In addition, all participants were presented with a series of forecasting questions similar to those they ultimately saw in the study, and practiced providing some forecasts and supporting rationales.

Delphi Forecasting Task for All Participants

The forecasting task used was based on an entrepreneurial class and competition called the New Venture Challenge (NVC) that is held annually at the University of Chicago's Booth School of Business. We had participants watch a short video presentation of each of three NVC teams, as well as review a 35-page packet of materials including information about each of the three NVC companies/teams. Based on their review of these materials, participants were asked to predict which of the three teams ultimately won the NVC that year.

Numeric Delphi (DN) Participants

After reviewing 3 video clips, one of each of the NVC teams, participants were asked to make their initial forecast as to which team was the winner. Participants in the DN condition completed a form indicating the following: (1) the NVC team they think won; (2) their confidence level in their prediction of the winner; and (3-7) scores out of 100 for each of the packets of materials they reviewed. At the end of the initial round participants were presented with an anonymized report showing what their group members predicted in terms of the same question. They also were provided a histogram showing the distribution of votes among their teammates. At the start of the second round, participants were asked to reconsider their initial answers based on the feedback they had been presented. Then, once again they answered the same questions and again got to see how their teammates voted. This was repeated for a third time with participants' third round answers being considered final.

Delphi with Reasons (DR) Participants

After reviewing 3 video clips of each of the NVC teams, participants were asked to make their initial forecast as to which team was the winner. Participants in the DR condition completed a form indicating the following: (1) the NVC team they thought won; (2) their confidence level in their judgment of the winner; (3-7) scores out of 100 for each of the packets of materials they reviewed; and (4-10) reasons or rationales for why they thought the team that won, had won. At the end of the initial round participants were be presented with an anonymized report showing what their fellow group members predicted in terms of the same. They also were provided a histogram showing the distribution of votes among their teammates, and—in this condition—a list of each reason or rationale shared by their teammates. At the start of the second round, participants were asked to reconsider their initial answers based on the feedback they had been presented. Then, once again they answered the same questions and again, got to see how their teammates voted, and reviewed their reasons or rationales. This was repeated for a third time with participants' third round answers being considered final.

Debrief for All Participants

Upon completion of the study (i.e., at the end of the third forecasting task), participants were asked to provide some basic demographic data including their gender, age, year in the MBA program, etc.

Study 1: Hypothesis

We originally hypothesized that Delphi with Reasons (DR) groups would outperform Numeric Delphi (DN) groups on appropriate measure of judgment accuracy. As such we were cautiously optimistic that our data would at least be suggestive of DR groups performing at a higher level when compared with DN groups on one or more of our proposed measures of accuracy.

Study 1: Analysis and Results

Unfortunately, after running our experimental test, our results were disappointing. We looked at various comparisons between conditions, but no matter which measures we utilized we found no statistically significant differences between the two Delphi Method conditions. We looked at a range of measures/comparisons: These included picking the winner; getting the exact order correct (i.e., correctly specifying which NVC team came first, which came second, and which came third). We adjusted for confidence ratings; we scaled the differences; we looked at improvements between rounds 1 and 2, rounds 1 and 3; and we tried using medians and means; all to no avail.

Recall the sample comprised 137 Booth MBA students enrolled in a Business Strategy course (73% male, 29% STEM college majors, 48% Business college majors, and 35% intending to pursue a finance specialty career). Individual participants were assigned to 20 groups with 6, 7, or 8 members. All analyses are based on "group" as the basic unit of analysis in the data set.

The primary dependent variables measuring quality of performance were based on the criterion of the ranking assigned to the NVC Proposals by a panel of expert business analysts and entrepreneurs as part of a competition sponsored by the Booth School. One measure was a simple dichotomous (0/1) match-the-winner score (the "winner" was the top-ranked proposal by the expert NVC Judges' panel of the 3 proposals presented to the MBA participants; each group received a score of 1 if the average individual rating on a 1-5 point scale indicated the expert's 'winner' was top-ranked for that group, 0 otherwise). The second measure was a confidence-weighted match score, essentially the confidence assigned to the "winner" averaged across all members of the group (confidence rated on a 1-5-point scale).

Overall average results are shown for the Match-the-Winner measure in Figures 1 and 2; an analysis of variance showed no significant effect of the manipulated Delphi Method treatment (Delphi with Reasons versus Delphi Numerical; F(1,18) = .19, n.s.), the effect of estimation rounds was significant (F(2,36) = 5.363, p < .01), with accuracy increasing substantially from Round 1 to Round 2 and leveling off between Round 2 and Round 3.

Figure 1. Box plot summary of the proportion of groups members (in each team) who picked the winner of the NVC competition, across rounds 1, 2, and 3 of judgments, comparing Delphi With Reasons versus the Delphi No Reasons teams.

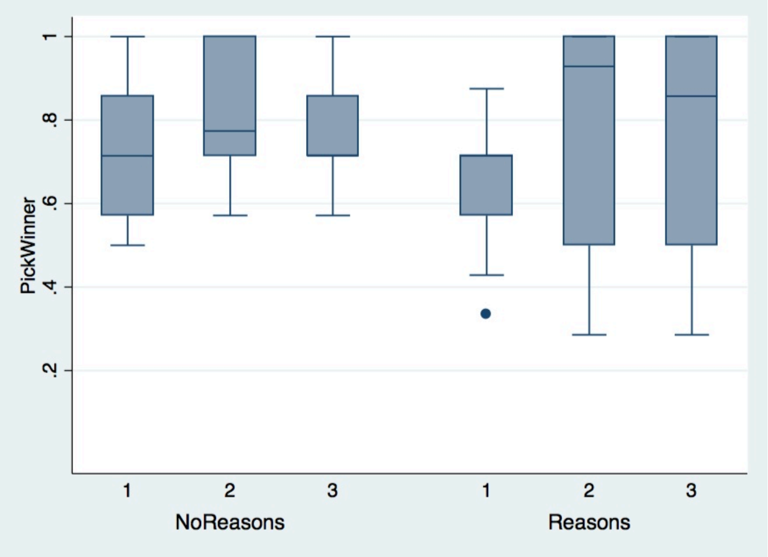
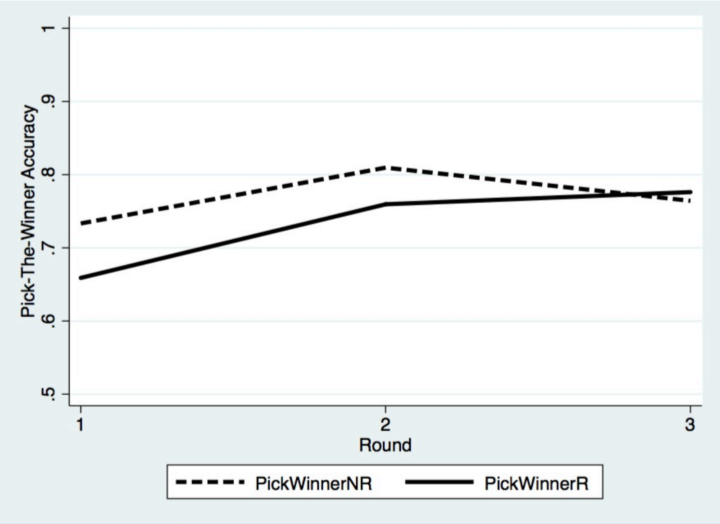


Figure 2. Line graph summary of the average proportions of group members (in each team) who picked the winner of the NVC competition, across rounds 1, 2, and 3 of judgments, comparing the Delphi With Reasons (solid line) versus Delphi No Reasons (broken line) teams.



The Confidence-Weighted-Accuracy measure showed the same pattern of results (Figures 3 and 4), and significance levels (F(1,18) = .04, n.s. for the comparison between Delphi Methods; F(2,36) = 10.71, p < .001) across the 3 rounds of judgments).

Figure 3. Box plot summary of the Confidence Adjusted Accuracy Scores of group members (in each team) who picked the winner of the NVC competition, across rounds 1, 2, and 3 of judgments, comparing the Delphi with Reasons versus the Delphi No Reasons teams.

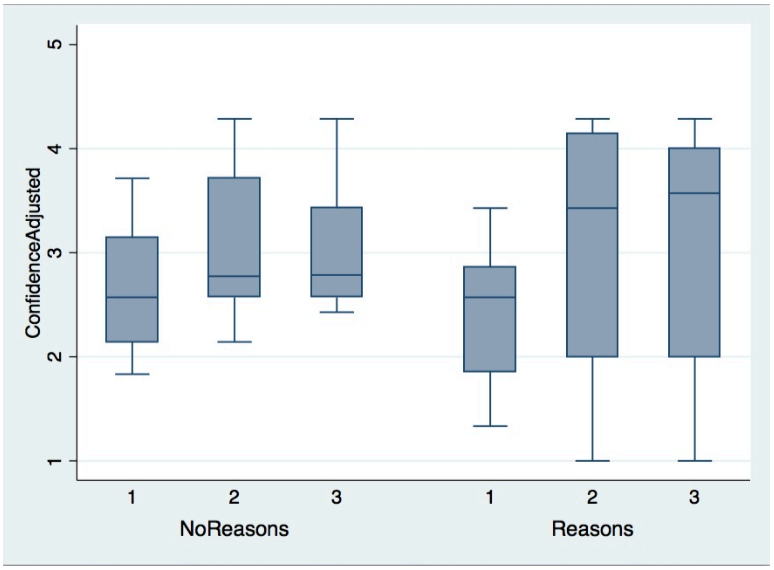
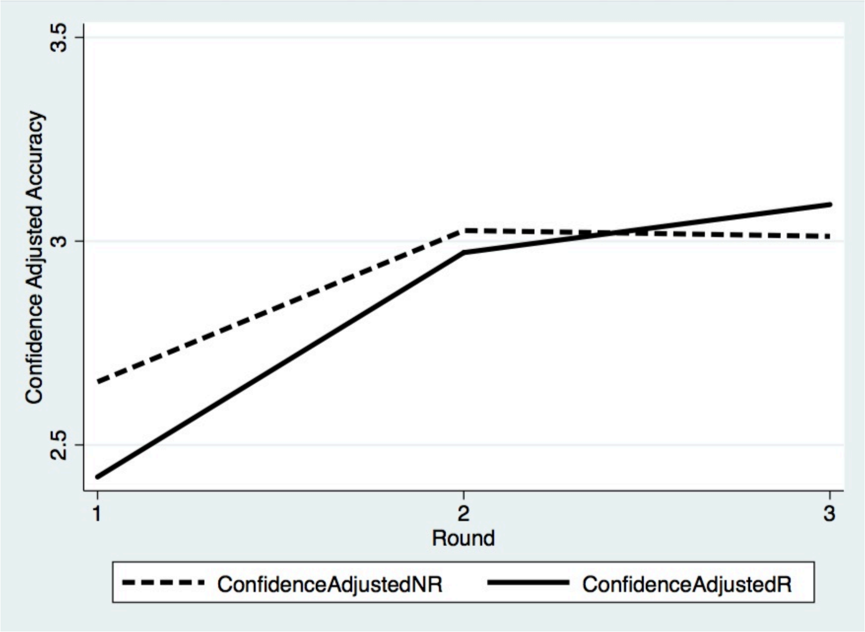


Figure 4. Line graph summary of the Confidence Adjusted Accuracy Scores of group members (in each team) who picked the winner of the NVC competition, across round 1, 2, and 3 of judgments, comparing the Delphi With Reasons (solid line) versus the Delphi No Reasons (broken line) teams.



Study 1: Discussion of the results

In study 1, we tested our initial ideas of how to increase performance by enhancing information pooling. As noted, analysis of variance showed no significant effects of the manipulated Delphi Method treatment. Although our initial reaction to these results was disappointment, upon further reflection we have come to the realization that perhaps we ought to have expected these results and not be surprised and disappointed. This is because of the nature of the New Venture Challenge judgment task we asked participants to perform in study 1. With reference to our earlier discussion of McGrath (1984), the tasks in study 1 were "intellective" and fell in the subset of (iii) because the "correct" answers were based on a consensus of experts, i.e., the judges of the New Venture Challenge.[10]

We find ourselves echoing the words of McGrath himself who writes: "Why, after all, should a group accept the nonobvious and not compellingly demonstrated answer of a minority of one, even though the investigator knows that that answer is correct?" And as he goes on: "The salience of this question will become even greater as we consider group performance on tasks for which the correct answer is defined by expert consensus."

In other words, our methods did not work, but perhaps we should not have expected them to work when the "truth" is not objectively true, is not intuitively compelling, is difficult to demonstrate, and is based on a consensus of experts rather than some objective and verifiable external measure. And perhaps it is not unreasonable to think such methods might be further handicapped when all participants consider themselves, rightly or wrongly, to be an expert in the field, and/or when all participants have almost identical information, i.e., no unshared information as would be the case in a hidden profile problem. In study 1 all participants received exactly the same information packets and viewed exactly the same video clips relating to teams competing in the NVC.

On balance, the take away from study 1 was that we came to recognize that our methods, if they were to work at all, would have to be tested in an environment (specifically on a set of tasks) even more amenable to information pooling than the start-up evaluation task. As a result of study 1, and utilizing the typology proposed by McGrath (1984), we decided our follow-up study should study tasks for which there is a demonstrably correct answer that is objectively

[10] It could be argued that in study 1 the tasks asked of participants were might be considered Decision-Making Tasks (per McGrath's typology), but I think that is not the case because participants are being asked to pick the winner knowing that the winner was defined as the New Venture Challenge team chosen by a panel of judges.

verifiable. In other words, any future study ought to use tasks that were still “intellective,” but in the subset of (ii) comprising tasks that have demonstrably correct answers even if, in each case, the correct answer is difficult to demonstrate.

Study 2: Controlled Experimental Test of Delphi with Enhanced Information Pooling

Based on the results from Study 1, further review of recent related research, refining our thoughts on the matter—with a weighty dose of intuition—we believe we now have a better handle on how to more effectively study information pooling. Rather than a classroom environment, in Study 2 we chose to conduct a controlled lab experiment. We also used questions that were likely to involve some unshared information and that could be better solved by teams in which members effectively pool their expertise and/or local knowledge.

Participants were asked to make a series of judgments about factual questions and intellectual puzzles – analogous to, for example, predicting the outcome of a sports event or a political election. An example factual almanac-type question is: “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?” The following is an example of an intellectual puzzle or brainteaser: “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.”

Participants were also asked to provide unshared, valid, judgment-relevant insights to better facilitate information pooling. Depending on the condition to which they were assigned, they might also have been asked to share their answers and insights with other participants. For

Study 2 we utilized an experimental design in a lab-based setting and sought to determine the relative accuracy of different information pooling techniques. We utilized three conditions: (i) Individuals Acting Alone (I); (ii) Basic Delphi Method (B); (iii) Enhanced Delphi Method (E).

Study 2: Participants

Participants were drawn from individuals recruited by the Center for Decision Research at the University of Chicago's Booth School of Business. The Center samples volunteer participants from two locations: the DRL (in the Harper Center on the main campus of the University of Chicago in the Hyde Park neighborhood of Chicago) and the CRL (in downtown Chicago). Participants were primarily undergraduate students from a mix of Chicago-area colleges including the University of Chicago.

We successfully recruited 222 participants who were assigned to 3-person groups. We had 74 groups. Each group was then assigned to 1 of the 3 experimental conditions: 23 groups to the Individual condition; 25 to the Basic condition, and 26 to the Enhanced condition. We used both locations with an approximate split of 50/50 between the two locations for questions 1 though 7, and, as the numbers above suggest, we had a good balance across conditions.

Scheduling participants worked slightly differently in the two locations: In the DRL, in Hyde Park, participants are almost all "walk-in" undergraduate students who come by the DRL looking to participate in a study. On the other hand, in the CRL in downtown Chicago, participants are scheduled ahead of time by email and arrived in groups of 3 randomly assigned to 1 of the 3 conditions. At the DRL, if a participant arrived and there were other participants also looking to participate in a study, then we formed them into groups of 3 and, on an alternating basis, assigned these groups of 3 to condition "B" or "E". For participants arriving at the DRL, when traffic was low, and/or a quiet day was expected, then we assigned such individuals to condition "I" and had them proceed through the study individually. In such cases, we still assigned a Group ID but recognized it is really a placeholder for future analysis. All participants were assigned a unique participant ID and a group ID.

Table 3. Demographics of Study 2 participants by location

Trait	**CRL**	**DRL**
Gender	31% Female/69% Male	47% Female/53% Male
Age	Mean = 34.85 (s.d. = 11.86)	Mean = 20.75 (s.d. = 2.20) =
Frequency of CTA Use	63% = "Every Day"	5% = "Every Day"
College GPA	Mean = 2.99 (s.d. = .788)	Mean = 3.44 (s.d. = .370)
SAT Score	Mean = 513.95 (s.d. = 666.66)	Mean = 2532.64 (s.d. = 2758.44)
Attended an Elite College?	4.96% = "Yes"	100% = "Yes"
Interested in Follow-up?	80% = "Yes"	73% = "Yes"
Lived in more than 1 region of USA?	27% = "Yes"	35% = "Yes"
Have only lived in the Mid West?	71% = "Yes"	55% = "Yes"

As can be seen from the background information in the table above, the participants recruited by the CRL were importantly different than those recruited by the DRL. The DRL participants were more gender balanced, younger, had higher GPAs and SAT scores, without exception attended elite colleges, and had lived in more regions of the USA than just the Midwest. CRL participants were almost 70% male, typically in their 30s or 40s, had average (or lower) GPA and SAT scores, attended colleges such as Roosevelt University, Moraine Valley Community College, Malcolm X College, Harold Washington, Robert Morris, and Rider University; and they often had lived only in the Midwest.

Study 2: Materials

The study was driven by a series of Google docs, specifically five forms (and associated spreadsheets which would contain the data submitted by the participants). Three of the forms were condition-specific, i.e., three of the five forms were each matched to the condition a participant has been assigned to. These three primary forms were called the “Individual Form” for condition I; the “Basic Delphi Form” for condition B; and the “Enhanced Delphi Form” for condition E. There were also two common forms that all participants would see at the beginning (“Training Form”) and at the end (“Debrief and Demographic Form).

All the forms were linked with a spreadsheet that collected the data that each participant entered into the form. The complication in this study comes from the fact that for two of the forms (“Basic Delphi Form” for condition B and “Enhanced Delphi Form” for condition E) we looked to share some of the data in the appropriate linked spreadsheet with participants at two occasions during the study (after participants completed their round 1 answers and insights and again after participants completed their round 2 answers and insights).

The data management in regard to the other forms is much more straightforward. While we used the “Training Form” to gather some information, as expected we found little of interest to analyze in the linked spreadsheet. And while we are interested in what information was captured in the linked spreadsheets for “Debrief and Demographic Form” and “Individual Form” because there is no sharing of that information with participants, the spreadsheets, could be left alone to accumulate data for the entire period in which the study is running.

The “Training Form” gives two example questions with example rationales or insights. The first is more of a factual or almanac question (but combines aspects of a brainteaser): “How many ping-pong or golf balls can fit inside a Boeing 747 airplane configured as a freighter?”

While the second is more of an intellectual puzzle or traditional brainteaser: “How many piano tuners are there in greater Chicago?” The possible answers we provided were 19 million and 200 respectively (the first is correct according to Boeing’s own website; the second is our own estimate for piano tuners in Chicago). The example rationales or insightss relating to the number of golf balls in a 747 we provided for training purposes were: “You might have a vague memory of a 747 tail height being equivalent to a six-story building, or that a golf- ball has a volume of approximately 2.5 cubic inches. Or you might have a sense of a 747 approximating the size of four average size single-family homes (at 1375 square feet each).” Similarly for the piano tuners questions, we provided the following example rationale(s) or insight(s): “You might know there are 2.5 million households in Chicago and that roughly one in twenty households has a piano. Or you might have a sense that pianos need to be tuned once every year. Or that it takes about 2 hours on average to tune a piano.” Participants were then asked to tackle their first question without guidance: “How many faculty members are there at the University of Chicago?”

Using the “Training Form” we provided guidance in terms of what a unshared, valid, judgment-relevant insight or rationale might look like, instructing participants that we would like them to share, “What insight(s) helped you arrive at your answer for the number of faculty members at the University of Chicago? 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?”

The forms used for the majority of data gathering each contain the same questions. Participants submitted the same condition-specific form three times, once for each of the three rounds in which they answered the questions and provided their insights. These three forms asked that participants answer 7 (or 10) questions that are a mix of almanac and brainteaser type questions, including a single Raven's IQ test question. For all questions and across all conditions, participants were asked to, "Please provide some insights that were the basis for your answer." The other 6 (or 9) questions included in these forms are displayed in Table 2 (below).

The final form, called the "Debrief and Demographic Form," asked each participant to provide some basic demographic data including gender, age, college, college graduation year, major field, and college GPA. It is our hope that some of these capture something akin to "expertise" as it relates to the questions asked in Study 2, more general expertise, and/or cognitive ability. We also asked participants to specify which regions of the US they had lived in and how frequently they rode on Chicago's CTA public transportation system. These last two were meant as proxies for specific kinds of "expertise" or "local knowledge" in relation to specific questions asked in the study. We also asked participants to provide their email address (should we need to contact them upon award of prize money) and that they express an interest (or not) in receiving answers to the questions and other study-related materials upon the conclusion of the study.

Table 4 provides a list of the estimation questions that composed the basic set of to-be-judged problems. Questions 1, 2, 3, 4, 6, 8, 9 and 10 all had objective numerical answers; and Questions 5 and 7 had binary correct/incorrect answers. Questions 1-7 were chosen because, intuitively they met two criteria that promote the value of information-pooling: (i) There is a broad range of relevant information, and (ii) there was reason to suppose that the information might be partly unshared among a diverse sample of group members. While Questions 8-10 were based on questions popular in other studies of group estimation, but intuitively they did not seem likely to show an advantage for information-pooling. The point of including these last questions is to see if any advantages of information-pooling (in the Enhanced Delphi condition) are diminished for these less information-rich questions.

Question 7 was a medium-difficulty test item from the Raven's Progressive Matrices nonverbal test of general intelligence (Figure 5). Test takers are asked to identify the missing element that completes a pattern, i.e., choose which of 8 images completes the pattern if used to fill in the lower right block. Many patterns are presented in the form of a 3x3 matrix. Easier questions usually have rules that apply both to columns and rows, so a test taker can choose to look at either to solve the puzzle; however harder puzzles either use only rows or only columns. In order to solve the problem, participants needed to recognize that the lines are rotating, one small segment at a time, and choose which of 8 images 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.'

Figure 5. Raven's Progressive Matrix used as Question 7

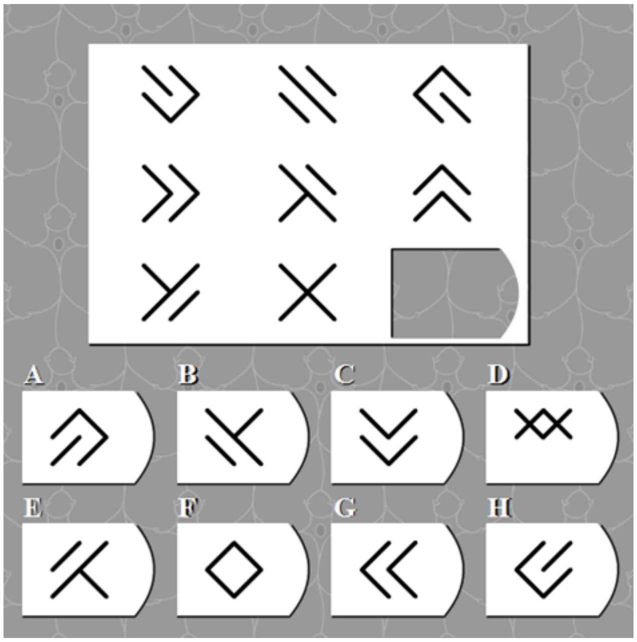


Table 4. Study 2 Questions 1 through 10 (excluding Raven's Question 7)

#	Question	Answer	Source/Explanation
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.
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.
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/
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
5	According to data from the 2010 U.S. Census's American Community Survey, which of the following U.S. metros 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;
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.

Table 4, continued

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
9	How many murders were officially registered in the US in 2013?	14,827	Source: FBI Crime Statistics.
10	In 1900, the percentage of the total US populations that was aged 65 and over was 4.1%. What was the percentage in 2013?	13.1%	Source: Administration on Aging.

Study 2: Procedure

Set-up Before Arrival of Participants

Before participants arrived at the DRL or CRL, our RAs ensured they were ready to provide all participants a consent form. On that consent form RAs were instructed to have written a unique Individual ID and unique Group ID (per a master list). RAs also ensured workstations were setup such that the appropriate forms were viewable from a full screen browser window, and that participants in condition B, at the appropriate time, were able to view the previous round answers from their two group-mates. Similarly, we ensured that participants in condition E, at the appropriate time, were able to view the previous round answers and insights from their two group-mates. It is also worth highlighting that the key distinction between the two Delphi groups, i.e., condition B and E (and the focus of our study), is the impact of sharing not only round-to-round answers between group-mates, but also round-to-round rationales and insights, so this part of the procedure was critical.

Data management was central to the set-up. RAs were asked to delete old data from the "Delphi-Basic Participant Sheet" and the "Delphi-Enhanced Participant Sheet." Except for the first time participants were run through the study, each of those spreadsheets contained 9 rows of data, 3 rows per participants, 1 row representing each participant's answers and insights per round. Those data were originally copied and pasted from the "Delphi-Basic Groups MASTER Sheet" and the "Delphi-Enhanced Groups MASTER Sheet" so could be deleted safely.

When the spreadsheets discussed above were created, we hid certain columns intentionally so that participants in condition B (Delphi-Basic) could provide insights but those insights were not shared among participants on a round-to-round basis (and nor could participants see their own insights from current or prior rounds). Given the centrality of this manipulation to our study, we asked that RAs double-check that participants in condition B were never shown columns pertaining to insights, but could only see columns containing answers to the questions.

All Participants Arrival and "Training Form"

RAs welcomed participants to the DRL or CRL and provided each of them with a consent form. As noted earlier, all participants in the CRL were scheduled to arrive in groups of 3. Whereas if participants arrived at the DRL and there were other participants also looking to participate in a study, then we attempted to find groups of 3 and, on an alternating basis, assign them to condition "B" or "E". If a participant arrived at the DRL, traffic was low, and/or a quiet day was expected, then we assigned such individuals to condition "I" and had them proceed through the study individually.

RAs then took each participant to an individual workstation displaying the first form, which was called "Training Form" and was the form used to train participants about how to answer the questions in the study and, in particular, to give them some specific instructions about what kinds of insights they ought to share with their group members (in the E condition). Again, all participants regardless of condition utilized this form. Once participants had read and submitted their responses to the "Training Form," they were directed to one of the three condition-specific forms through which we gathered the primary data for our study. Again, these three principal forms are named the "Individual Form"; the "Basic Delphi Form"; and the "Enhanced Delphi Form."

Participants in Individual Condition (Rounds 1, 2, and 3)

Things were pretty straightforward for individuals assigned to condition I. They submitted answers and insights to their condition-specific form 3 times (i.e., they submitted the very same form 3 times with a short break between each round to consider their previous round answers). We anticipated some participants might express frustration at this procedure and, as a result, asked that our RAs, if faced with such objections, simply to remind participants (as it says on the form) that there is empirical evidence to suggest re-considering prior answers is beneficial in certain circumstances.

Participants in Conditions B and E (Round 1)

For the participants in groups of 3 and assigned to conditions B or E, this is where things got somewhat complicated. Participants in condition B and E were shown a limited set of shared group-specific and condition-specific data from the underlying spreadsheet. To ensure data and experimental integrity, we asked that RAs not share the master spreadsheet underlying the form in question, and instead only to share the appropriate participant sheet (which per procedure they would have cleaned up prior to the arrival of the current participants).

As participants in condition B and E completed the "Basic Delphi Form" for condition B and "Enhanced Delphi Form" for condition E, from another workstation RAs watched the appropriate master spreadsheet live (i.e., while it updated with information as the actual groups of 3 participants submitted their first round answers and insights).

When all 3 participants in a group completed the form, RAs then copied and pasted the relevant rows into the appropriate participant spreadsheet and showed this spreadsheet to the participants. RAs accomplished this by loading the appropriate spreadsheet on each of the 3 participants' workstations. In the first round this form would have 3 rows of data in it (i.e., one row for each participant); in the second round this same form would then have had 6 rows of data in it (i.e., two rows for each participant). We instructed RAs that there is no need to delete data from the participant spreadsheets between rounds 1, 2, and 3, as we were comfortable with all previously input data being viewable by participants, not just what was input in the previous round.

We let participants review the spreadsheet for approximately 2 minutes. Then our RAs closed the spreadsheets on the participants' workstations, opened up the appropriate condition-specific form once again, and requested that participants complete and submit the form a second time.

Participants in Conditions B and E (Round 2)

Again, from another workstation RAs watched the appropriate master spreadsheet live (i.e., while it updated as groups of 3 participants submitted their second round answers and insights). When all 3 participants in a group completed the form a second time, our RAs copied and pasted the relevant rows from the master spreadsheet into the appropriate participant spreadsheet and showed this spreadsheet to all 3 participants on their workstations. The spreadsheet then had 6 rows of data in it (i.e., two rows for each participant) because there was no need to delete data from the participant spreadsheets between rounds 1 and 2.

For a second time, RAs let participants review the spreadsheet for approximately 2 minutes. Then our RAs closed the spreadsheets on the participants' workstations, opened up the appropriate condition-specific form once again, and requested that participants complete and submit the form a third and final time.

Participants in Conditions B and E (Round 3)

Although participants answered the questions and submitted the form a third time, there is no sharing of information between group members subsequent to the third and final round. Given this, once participants submitted their form a third time, they were considered ready to be debriefed and to answer a few demographic questions.

All Participants “Debrief and Demographic Form”

Once participants, regardless of their condition, submitted their final third round answers, our RAs directed them to the last form they were asked to complete. All participants regardless of their assigned condition were asked to complete the “Debrief and Demographics Form.” In it, we asked that participants to provide some basic demographic data including gender, age, college graduation date and academic field(s) of study, if they attended college. We also asked their email address (should we need to contact them upon award of prize money) and asked them to express an interest (or not) in receiving answers to the questions and other study-related materials upon the conclusion of the study.

Study 2: Measures of Accuracy

There are obviously all manner of ways in which we might measure accuracy. Graefe and Armstrong (2011) chose to use the mean absolute error (MAE) to measure how close forecasts or predictions were to their eventual outcomes. As the name suggests, the mean absolute error is an average of the absolute errors, averaged across members within each group. MAE is just one of several alternative ways of comparing forecasts with their eventual outcomes. Other well-established alternatives include the mean absolute scaled error (MASE) and the mean squared error. These all measures summarize performance in ways that disregard the direction of over- or under-prediction. In contrast, a measure that does place emphasis on over- or under-prediction is the mean signed difference.

We choose to use what has become the accepted norm for research in this and related fields, The Percent Absolute Error score: take the absolute value of “Estimate” minus “Truth” divide by the “Truth” and multiply by 100 (e.g., Mellers et al., 2014; Gürçay, 2014). For example, Question 1 from Study 2 asks participants the height in feet of the highest man-made building. The tallest man-made structure as of May 2014 is the Burj Khalifa in Dubai, United Arab Emirates that stands 2722 feet tall. As such, a participant who answered 2722 would be perfectly accurate, i.e., using our preferred measure would have an error score of zero. Suppose an individual estimated the height of the tallest structure as 2600, her PAE score would be: [|2600 – 2722|/2722]*100 = 4.48

Our principal focus is accuracy in the final round. Table 5 shows how these measures are calculated, given various dummy data relating to potential estimates to the question of how tall in feet is the highest man-made building.

Why do we use such a measure? Well for one reason we feel that it best captures the common sense idea of accuracy and best represents what it is that individuals look to maximize if asked to give their most accurate answer to a question. In other words: if participants are asked to maximize absolute accuracy, they strive to get as close as possible to the metaphorical bull's-eye. Given this, we favor keeping things simple and measuring the absolute distance from the metaphorical bull's eye.

Accordingly, we will calculate measures of accuracy for all questions and for all participants, for all rounds, and for all questions. Once we have done so, we will then make between-condition comparisons to ascertain group performance for each condition.[11] For participants in actual groups, we will calculate group medians and means of these measures. For participants in individual conditions (only applicable to study 2) we will calculate group medians

[11] Gigone and Hastie (1997) suggested we assess group performance by treating the aggregated group performance as if it were that of a hypothetical individual and seeing where that hypothetical individual stacks up in terms of performance rank when compared with actual individual participants. We did not follow this advice, as with 3-member groups, the results were uninformative.

and geometric means based on statisticized groups of three, as well as comparing the group medians and means of all individual participants. We choose to use both group medians and means because since Galton, when it comes to group decision-making, there has been a marked preference for using medians (and medians obviously dampen extreme outliers which is likely a positive here). But we see no reason not to also make appropriate comparisons using means.

Table 5. Measures of Accuracy (based on dummy data)

	A	B	C	D	E
1	“Truth”	R1 Estimate	R3 Estimate	R1 Accuracy	R3 Accuracy
2	2722	2600	2600	4.48	4.48
3	2722	2722	2722	0.00	0.00
4	2722	2822	2622	3.67	3.67
5	2722	2500	2700	8.16	0.81
6	2722	2700	2900	0.81	6.54
7	2722	6000	1500	120.43	44.89

Although statistical analysis was completed in Stata, we choose to use Excel for initial data preparation and retrieval from the various Google documents. Should it not be clear, the values for Column D of Table 2 “R1 Accuracy” is calculated using the Excel Absolute (ABS) function which returns the absolute number, i.e, ignores negative/positive sign. So for the first row of data (Row 2) in Table 3, which relates to a single hypothetical participant, the value =ABS(B2-A2)/B2*100. Column E is calculated in much the same manner as Column D. These are error scores so the closer to zero the value in question, the more accurate the estimate.

We calculated accuracy measures for all questions and for all participants and groups. For all CRL participants and Delphi participants (condition B or E) who completed the study in the DRL, we then calculated group means and medians. For DRL participants in the individual condition (I) we calculated group medians and means based on statisticized groups 3 which were created after the fact. We also calculated means and medians of all individual participants considered as a whole. We choose to use both group geometric means and medians because these metrics damp the impact of extreme outliers on the analysis that is a benefit in distributions like the numerical estimates in this experiment.

Study 2: Hypothesis

Our primary hypothesis is that groups using the Enhanced Delphi Method (E) would make more accurate forecasts than those groups utilizing the Basic Delphi Method (B), and will make more accurate forecasts than Individuals Acting Alone (I) for questions 1 through 7. We have different expectations for questions 8, 9, and 10 than we have for the prior questions. Each of these questions was taken either verbatim or with slight modifications from other published studies (e.g., date was updated, geographic focus was amended to the US versus Europe).

Gürçay (2014) asked "What percentage of Americans has a pet?" and we used her question without further revision as our question 8. Lorenz et al. (2011) asked "How many murders were officially registered in Switzerland in 2006?" We took his question, updated the date to 2013, and changed Switzerland to US, given our study would run in the US not Europe. Our question 9 therefore read: " How many murders were officially registered in the US in 2013?" Graefe (2010) and Graefe and Armstrong (2010) asked participants "In 1900, the percentage of the total US population that was aged 65 and over was 4.1%. What was the percentage in 2000?" We took their question and simply updated the date such that our question 10 read: "In 1900, the percentage of the total US population that was aged 65 and over was 4.1%. What was the percentage in 2010?"

We do not think an enhanced information pooling process will improve performance in regards to questions 8, 9, and 10. This is because each of them asks for an obscure fact and it seemed to us, that in the case of obscure facts, information pooling will have little opportunity to improve a group's performance. Hence, we thought that enhancing the information pooling process would produce no discernable benefits on questions 8, 9, and 10.

Study 2: Analysis and Results

Our focus in Study 2 is on comparisons between social processes in group judgment tasks that promote information pooling versus processes that minimize pooling and maximize independence. Participants answered 10 estimation questions, either independently, with no information sharing, or in a Basic Delphi Method in which solutions were shared before estimations were made on later rounds, or in an Enhanced Delphi Method in which participants were instructed to share judgment relevant facts, opinions, and solutions between estimation rounds. (Recall that 3 questions, Q8-Q10, were selected because we expected that information-pooling would not improve accuracy on those questions).

In light of our earlier discussion, our dependent variable was "Percent Absolute Error" or PAE for questions 1, 2, 3, 4, 6, 8, 9, and 10. Questions 5 and 7 were scored in a binary manner: 1 indicating "wrong" and 0 indicating "correct." Note in all cases, zero is a perfectly accurate score. Our focus was on accuracy in the final, i.e., 3rd, round. For questions 1, 2, 3, 4, 6, 8, 9, and 10, zero is the perfect accuracy score and the higher the accuracy number, the less accurate the estimate. For questions 5 and 7, the quality of an estimate was binary, i.e., a value of 0 indicates "correct" while a 1 indicates "incorrect."

We examined the distributions of the Percent Absolute Error scores for each group across all three rounds and the questions (excluding questions 5 and 7). Distributions were skewed, so the principal statistical techniques we used were non-parametric. But, we wanted to use 3-member groups as the basic unit of analysis for descriptive and inferential statistics and so a decision had to be made about what function to apply to the 3 individual judgments to produce the “group estimate.” (In addition, note that we created 3-member nominal groups of participants for the Individual judgment condition.)

One solution that we adopted is to use the 3-member group median (the solution favored by Galton in the original papers on ‘Wisdom of Crowds’; Galton, 1907a, 1907b) of the Percent Absolute Error measure. Obviously, this will damp out many of the effects of outliers. We also conducted parallel analyses using the geometric mean of the 3-members’ estimates, to provide converging evidence on the reliability of our analyses.

Geometric means are useful when comparing different items where one goal is finding a single "figure of merit" for the items. Geometric means can be used when each item has multiple properties that have different numeric ranges. Using a geometric mean "normalizes" the ranges being averaged so that no range dominates the summary weighting. This is especially helpful in terms of analyzing data like our own where the majority of estimates are low, but a minority of outliers are found in the far right tail of the right-skewed distribution.

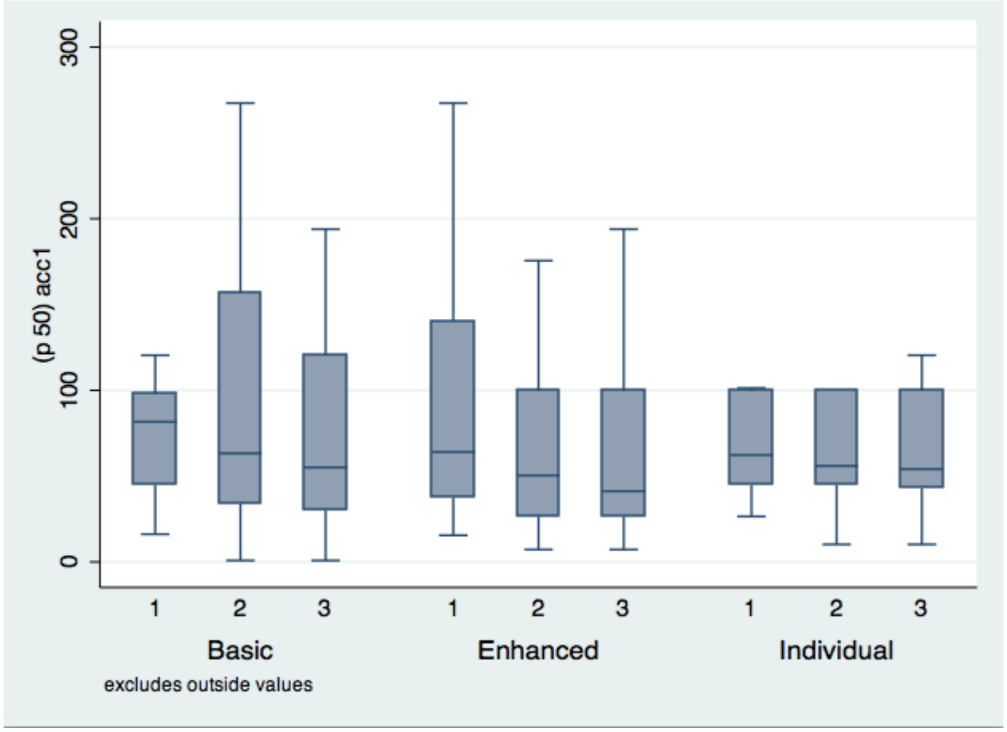
Our study was run in two locations, one (DRL) in which participants were typically students at an elite university (University of Chicago) and highly motivated; while in the other location, participants were recruited “off the street” (the CRL, in downtown Chicago) and were likely to be less expert, and less engaged in the task.

Our results are most promising among the participants sampled from the University of Chicago in Hyde Park (DRL) and less so among those participants who completed the study in the downtown location (CRL). The basic problem is the large number of outlier estimates, especially among participants who completed the study at the CRL. This is typical for this kind of study, cf. Gürçay, et al., 2014; Lorenz, et al., 2011).

Analysis of Judgments for Question 1: How tall is the highest man-made building in feet? (answer: 2722 ft., Burj Khalifa in Dubai)

A useful way to get an overview of the results is with a graphic. Here the challenge is the outliers that make displays uninformative by 'squeezing' most of the data into a small space in order to display the full range including the outliers. An informative display of over 95% of the data is provided by Box Plots that exclude outliers, with outliers defined as numbers higher or lower than 2.5 of the relevant "percentage Quartile" (observations "outside the whiskers" in a Tukey Box Plot). This would approximately exclude values +/- 3 standard deviations, if the data were normally distributed (which it was not). This graph (Figure 6) shows an orderly picture of the results, with errors decreasing smoothly across rounds for each condition, and Enhanced Delphi most accurate, followed by Basic Delphi, and then Individual group judgment conditions.

Figure 6. Trimmed box plots of Percent Absolute Error data for Delphi and Individual group methods; by round of judgments (excluding, approximately, the top and bottom 2% of the outlying observations).



The outliers themselves are meaningful, and are of practical importance. If a group judgment method is effective at reducing outliers, and the basic estimates are informative, then it has practical value. (Of course, we are hoping that our favored group judgment method, Enhanced Delphi, will do more than reduce outliers.) We tabulated extreme outliers, i.e., estimates where the Percent Absolute Error score was greater than 1000 on round 3 estimates (meaning the error was greater than 10 times the truth). We found, to our surprise, that extreme outliers occurred almost as often for Delphi Method judgments (whether Basic or Enhanced) as for Individual judgments (10 versus 7 outliers across both locations). But, as expected, more extreme outliers were observed in data from CRL participants, compared to DRL participants: CRL outliers on 16/105 (15%) Round 3 trials versus 1/117 (1%) for the DRL (n.b. this tabulation includes all 3 judgment rounds). This is not surprising as we did not force participants to follow the Delphi requirement to make next round estimates fall into the range from the prior round, yet the DRL participants almost always followed our instructions, while the CRL participants were more likely to deviate.

Table 6. Outlier Estimates for Question 1 (CRL).

Condition	**Number of Outliers**	**Number of Estimates**
Basic	0	33
Enhanced	9	39
Individual	7	33
Total	16	105

Table 7. Outlier Estimates for Question 1 (DRL).

Condition	**Number of Outliers**	**Number of Estimates**
Basic	1	42
Enhanced	0	39
Individual	0	36
Total	1	117

If we focus on the DRL participants alone, and look at round 3 accuracy in answering question 1, the median accuracy scores suggest the Enhanced Delphi Method is the winner.

Table 8. Median Round 3 Accuracy for Question 1 (DRL).

Condition	Median Round 3 Accuracy for Question 1
Basic	50.41
Enhanced	28.36
Individual	45.81

In terms of assessing significance, we took a conservative approach in light of our data quality issues. For statistical analyses we did not throw out data and instead relied upon common statistical techniques for addressing non-normally distributed data sets like our own. Such techniques allow us to sensibly “tune” our analyses in light of estimates that vary a lot, as is the case with our data (especially from the CRL location). Across most questions (those for which the accuracy score was not simply a binary zero or one, i.e., Questions 5 and 7), we took two approaches: (1) we looked at 'ranked ordinal data' using a Kruskal-Wallis test; and (2) we compared geometric means of our accuracy measures.

The Kruskal–Wallis one-way analysis of ranked data is a non-parametric method for testing whether samples – the three experimental judgment method groups – originate from the same distribution, so is well suited to analysis of data from our current study. Throughout we report the test with a correction for matching judgments (i.e., cases where two groups had identical median or geometric mean scores, and could not be ‘ranked’); although there are few pairwise matching scores in our data and the results never change depending on this correction. (The parametric equivalent of the Kruskal-Wallis test is the one-way analysis of variance (ANOVA) that assumes a normal distribution of the residuals.)

Table 9. Kruskal-Wallis analysis of ordinal ranks only for round 3 estimates of question 1 (CRL and DRL).

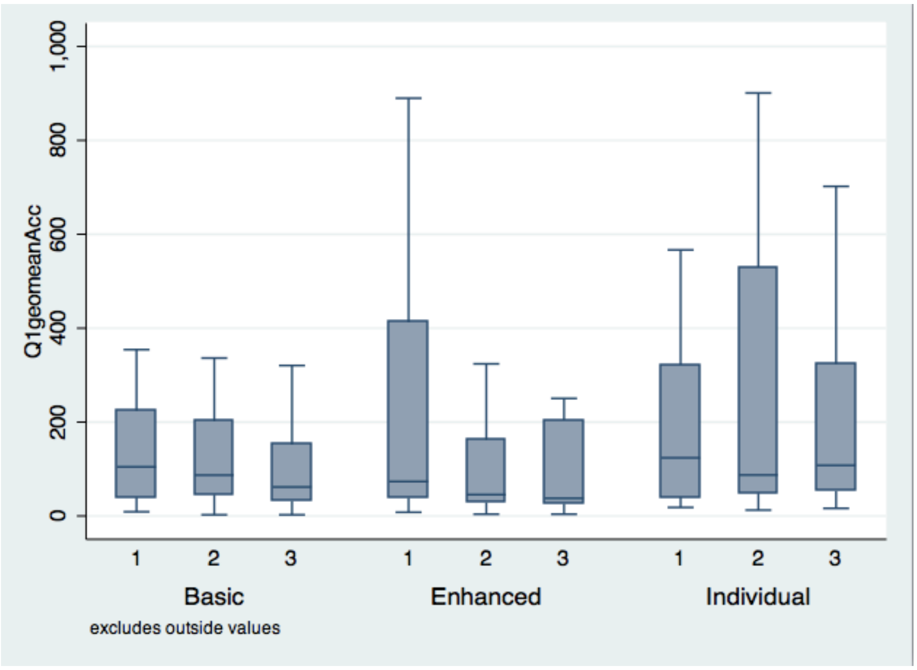
Condition	Observations	Rank Sum
Basic	25	957.50
Enhanced	26	877.00
Individual	23	940.50

The results in Table 9 suggest that the Enhanced Delphi condition is best (lowest rank-sum for the Percent Absolute Error score is most accurate), followed by the individual condition, and then with Basic Delphi bringing up the rear. There is a suggestion that Enhanced Delphi is most accurate, but all significance tests comparing conditions were non-significant (K-W Chi-squared (2 df) = 1.41, n.s.).

We repeated the Kruskal-Wallis analysis separately for the DRL and the CRL locations. As we expected the results are stronger and clearer for the DRL location (highly conscientious participants), versus the CRL location. Although again, non-significant (for the DRL results, K-W Chi-squared [2 df] = 5.31, p < .10; for the CRL, K-W Chi-squared = 0.58, n.s.)

We repeated all these analyses and reached exactly the same conclusions using the geometric mean of the individual judgments (rather than the median), within each 3-member group on each round. Results on this slightly different metric, were virtually identical to those obtained with the group median metric. Again the results were much stronger for the DRL location, in fact, reaching significance for that sub-set of the data: K-W Chi-squared (2 df) = 6.90, p < .05. Here (Figure 7) is a plot of the trimmed Box Plot data, for the geometric mean analysis; obviously the same general picture as obtained from the analysis on group medians:

Figure 7. Trimmed box plot summary of the values for geometric mean of the 3-member groups' Percent Absolute Error scores for Q1 (with outliers, "outside the Whiskers" excluded).



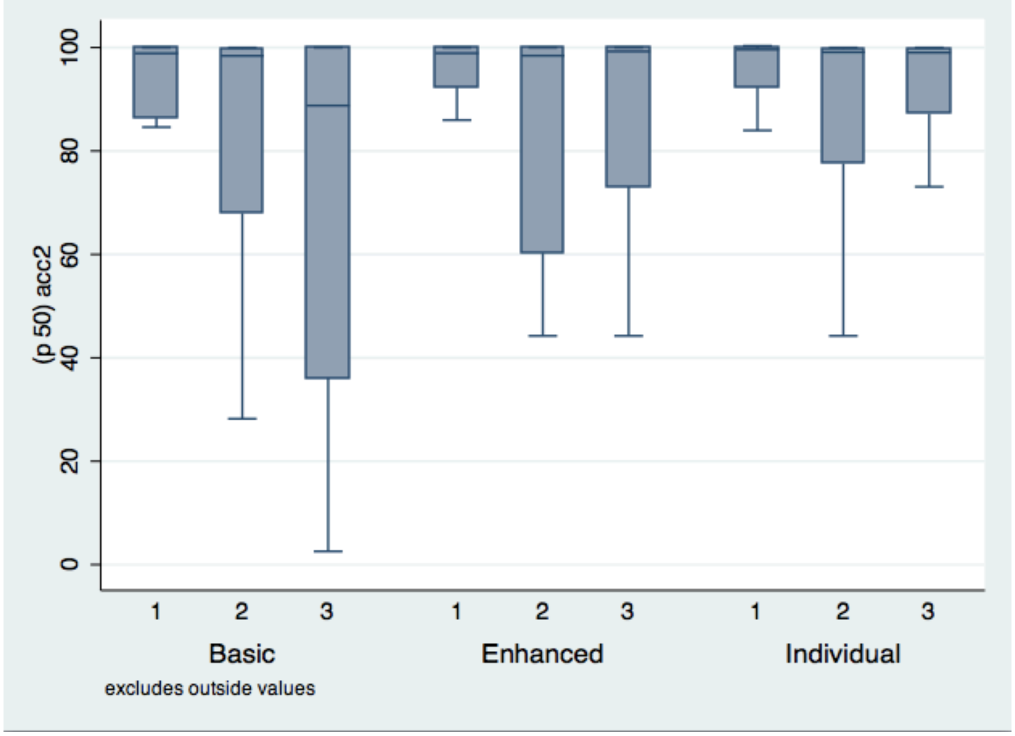
Analysis of Judgments for Question 2: What is the total area of the contiguous U.S. 48 states in square miles? (answer: 3,119,885 sq. mi.)

Question 2 became an important methodological lesson. The box plot summary of the data showed a bizarre pattern of apparently low estimates, with a massive peak in Percent Absolute Error scores near 100 (Figure 8). This led to an insight about the nature of the Percent Absolute Error metric. The insight is that there are always two values that yield a score of 100: (i) if the participant answers with a number near 0, as (|0 – TRUTH|/TRUTH)*100 = 100; or (ii) with an answer near 2 * TRUTH, as (|2*TRUTH – TRUTH|/TRUTH)*100 = 100. What we discovered was that many answers were in the range less than 100, a truly absurd value for "Total area of the 48 US states in square miles" (see Table 10). What participants were doing, in good faith, was submitting answers in the millions (e.g., 3.1 million would be a very good answer). The experimental procedure failed to constrain the responses to appear on a true square miles metric (where 3,100,000 would be a good answer), and so the answers to this question are not interpretable.

Table 10. Ranges of Numerical Answers to Q2.

Range	Frequency	Percent	Cumulative %
0-10	20	3.52	3.52
11-100	95	16.73	20.25
< 100	107	79.75	100.00

Figure 8. Trimmed box plot summary of the values for median accuracy for Q2 with outliers excluded.



Analysis of Judgments for Question 3: As of Financial Year 2012, how many Starbucks stores were there in the U.S.? (answer: 10, 924)

Question 3 was straightforward. The trimmed box plots (Figure 9) show the usual progression of improvement across rounds of judgments, and Enhanced Delphi is the winner as in Question 1. In addition, our expectation that Individual estimates would exhibit more variability across group judgment methods is confirmed more clearly in the box plots for Question 3, compared to the Question 1 data. However, the counts of extreme values (Percent Absolute Error > 1,000) did not show any systematic pattern; the counts were not greater for the CRL data (7/105) versus DRL (10/117), or across judgment methods (Basic, 5/70; Enhanced, 6/72; Individual, 6/63).

Figure 9. Trimmed box plot summary of the values for median accuracy for Q3 with outliers excluded.

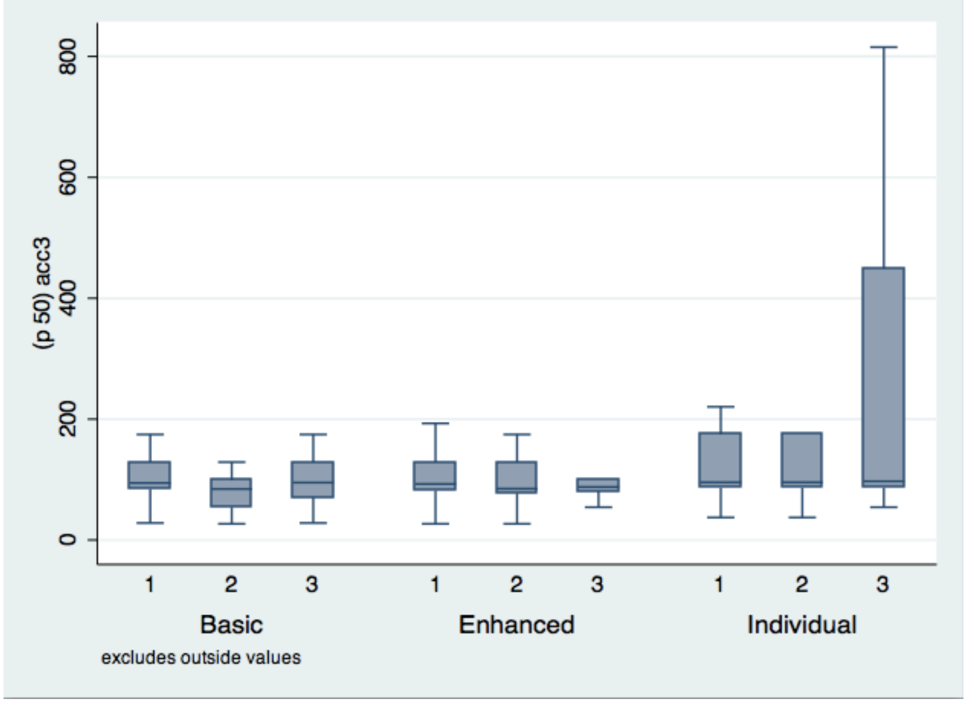
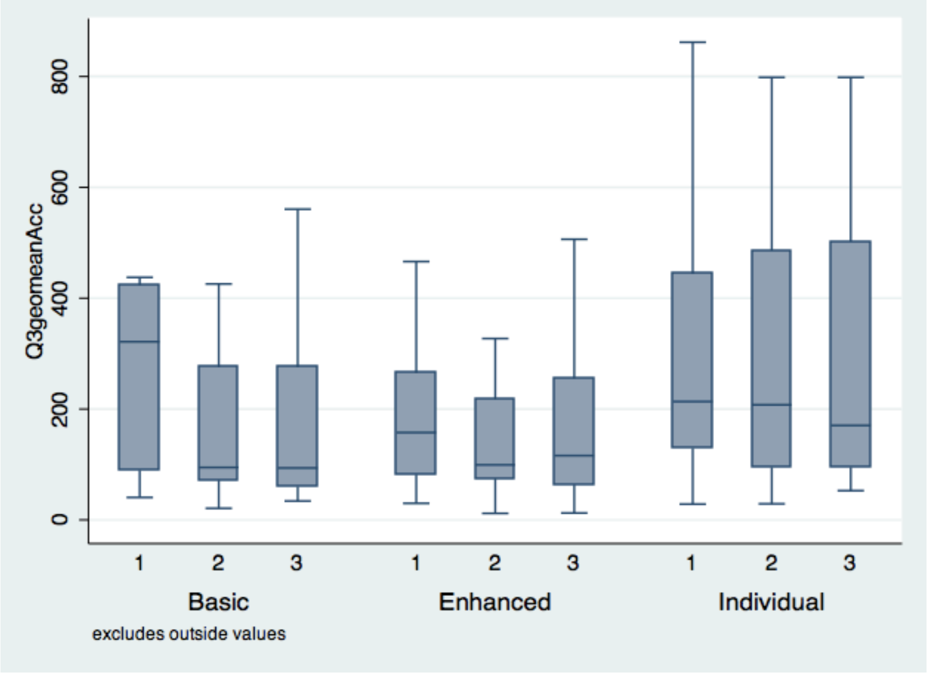


Figure 10. Trimmed box plot summary of the values for geometric mean accuracy for Q3 with outliers excluded.



The Kruskal-Wallis analysis suggests that the Individual condition estimates are less accurate (high rank sum for errors) than the Delphi conditions; but the effects are not significant (K-W Chi-squared (2 df) = 4.10, n.s.). Results from the DRL and the CRL, calculated separately also showed the same ordering from both locations (Individual considerably less accurate than the two Delphi conditions), but no results were significant (DRL: K-W Chi-squared (2 df) = 4.21, n.s.; CRL: K-W Chi-squared (2 df) = 3.27, n.s.).

Table 11. Kruskal-Wallis one-way analysis of ordinal ranks only for round 3 estimates of question 3 (CRL and DRL).

Condition	Observations	Rank Sum
Basic	25	866.50
Enhanced	26	873.50
Individual	23	1035.00

Analysis of Judgments for Question 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? (answer: 145)

The trimmed box plots (Figure 11) for Question 4 show the usual progression of improvement across rounds of judgments, although the pattern is not perfectly consistent. Enhanced Delphi is the winner on accuracy metrics as in Question 1. Here there is no obvious difference in variability across judgment methods. Counts of extreme values found no values in the extreme range (Percent Absolute Accuracy > 1,000).

Figure 11. Trimmed box plot summary of the values for median accuracy for Q4 with outliers excluded.

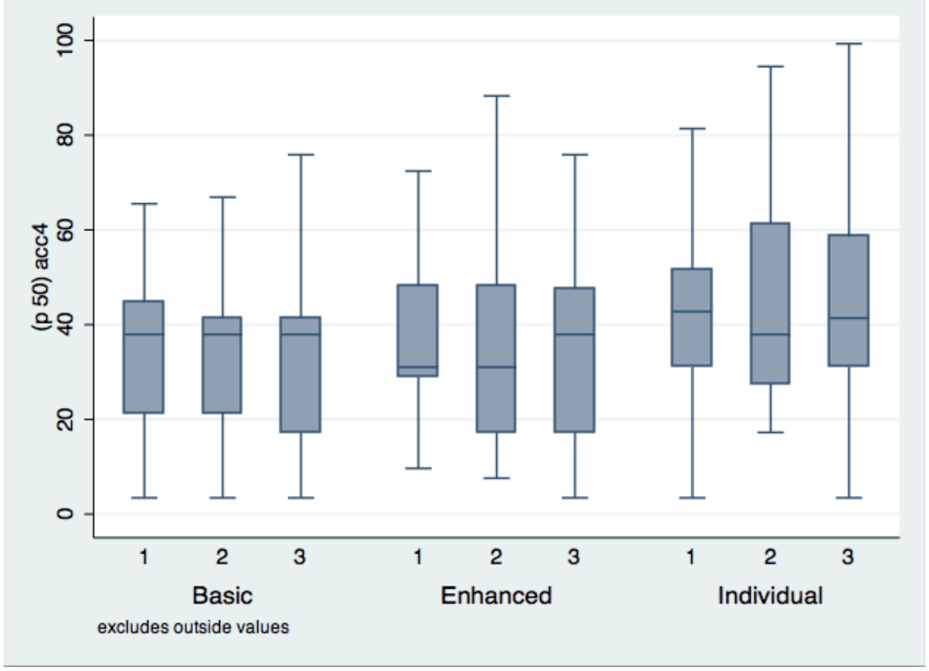
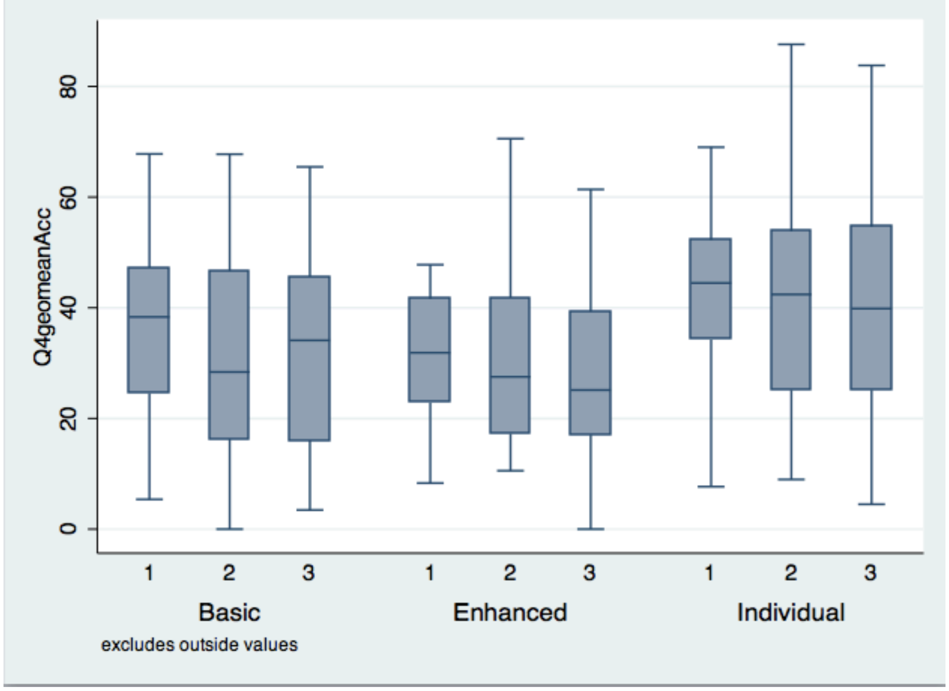


Figure 12. Trimmed box plot summary of the values for geometric mean accuracy for Q4 with outliers excluded.



The Kruskal-Wallis analysis suggests that the Individual condition estimates are less accurate (high rank sum for errors) than the Delphi conditions; but the effects are not significant (K-W Chi-squared (2 df) = 3.55, n.s.). Results from the DRL and the CRL, calculated separately also showed some differences between locations. The results from the DRL show the typical advantage for the Enhanced Delphi method, at marginal levels of significance (DRL: K-W Chi-squared (2 df) = 4.75, p < .10). The pattern from the CRL, found the Basic Delphi was most accurate, though far from significant (K-W Chi-squared (2 df) = 0.71, n.s.).

Table 12. Kruskal-Wallis one-way analysis of ordinal ranks only for round 3 estimates of question 4 (CRL and DRL).

Condition	Observations	Rank Sum
Basic	25	836.50
Enhanced	26	917.50
Individual	23	1021.00

Analysis of Judgments for Question 5: According to data from the 2010 U.S. Census's American Community Survey, which of the following U.S. metros has the largest percentage of foreign-born people: Miami, San Jose, Los Angeles, New York, Chicago? (answer: Miami)

Question 5 is the first with a binary answer (correct = 0, incorrect = 1), scored "in the same direction" as the numerical items. One summary comment on Question 5 is that it was very difficult, with error rates over 80% in all conditions. On round 3 the Proportion of Errors scores were: 24/25, 21/26, and 21/23 for the Basic Delphi, Enhanced Delphi, and Individual group judgment conditions respectively (non-significant with a standard Chi-squared test on the proportions (2 df) = 3.22, n.s.

If we separate the results by location, we see a hint of the Enhanced Delphi advantage for the DRL data: 14/14, 10/13, and 12/12 errors for Basic, Enhanced, and Individual, Chi-squared (2 df) = 6.50, p < .05. And, nothing in the CRL data: 10/11, 11/13, and 9/11 errors for Basic, Enhanced, and Individual, Chi-squared (2 df) = 0.39, n.s.

Analysis of Judgments for Question 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? (answer: 83 hours)

Question 6 was straightforward. The trimmed box plots (Figure 13) show the usual progression of improvement across rounds of judgments, and Enhanced Delphi is the clear winner. Here there is a surprise in that variability seemed higher across groups in the Delphi conditions, compared to the Individual condition. The counts of extreme values (Percent Absolute Error > 1,000) showed no clear pattern (2/75, 1/78, and 2/69).

Figure 13. Trimmed box plot summary of the values for median accuracy for Q6 with outliers excluded.

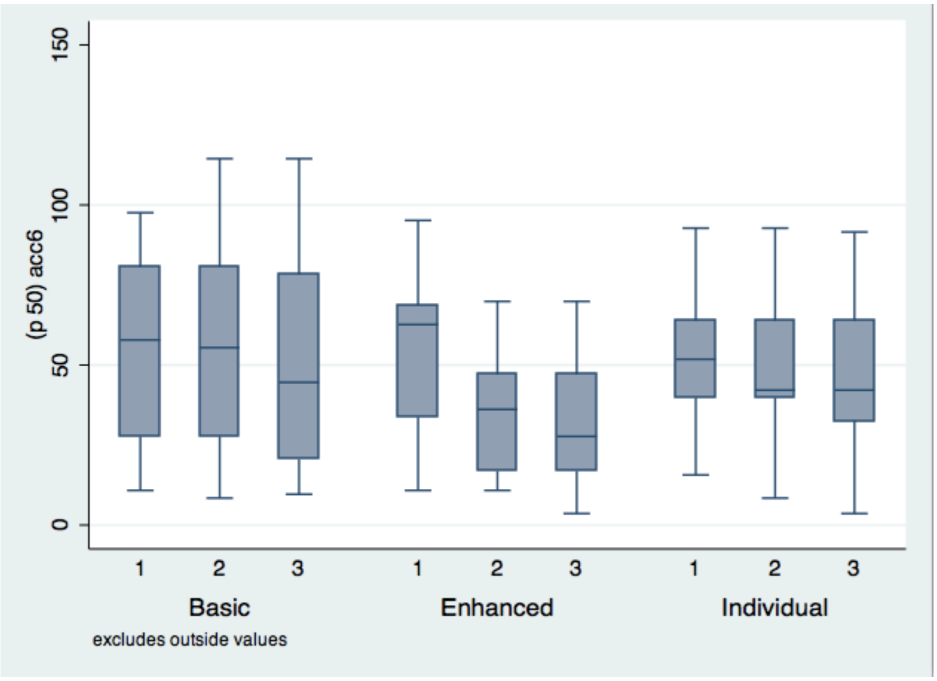
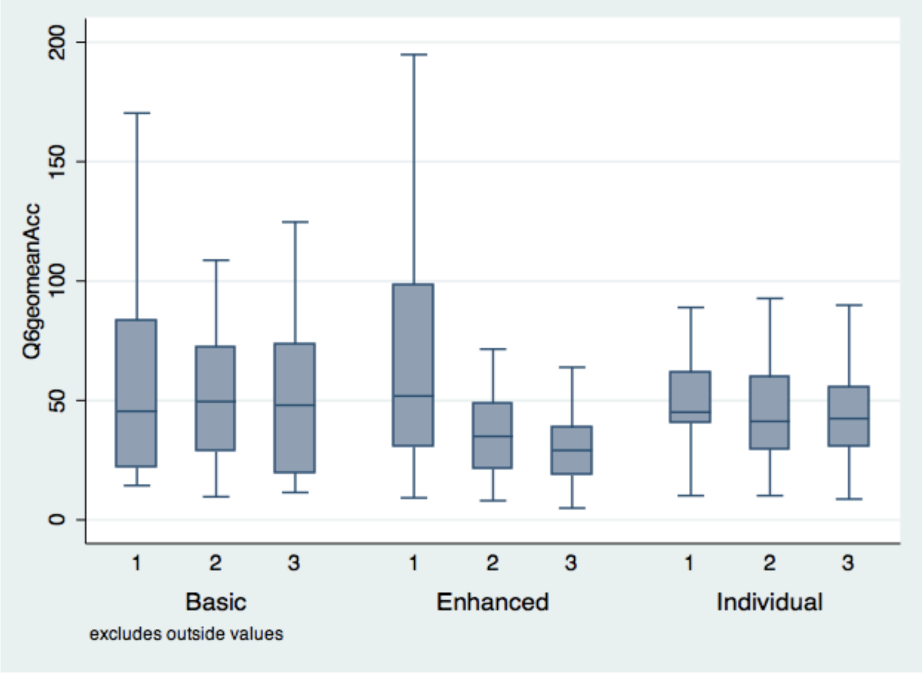


Figure 14. Trimmed box plot summary of the values for geometric mean accuracy for Q6 with outliers excluded.



The Kruskal-Wallis analysis suggests that the Enhanced Delphi is most accurate; but the effects are not significant (K-W Chi-squared (2 df) = 4.34, n.s.). Results from the DRL and the CRL, calculated separately showed the same basic pattern for both locations, though neither was significant (DRL: K-W Chi-squared (2 df) = 4.17, n.s.; CRL: K-W Chi-squared (2 df) = 2.65, n.s.).

Table 13. Kruskal-Wallis one-way analysis of ordinal ranks only for round 3 estimates of question 6 (CRL and DRL).

Condition	Observations	Rank Sum
Basic	25	1038.50
Enhanced	26	791.50
Individual	23	945.00

Analysis of Judgments for Question 7: Raven Progressive Matrices intelligence test (see Methods Section above for details)

Question 7 was also a binary (correct = 0, incorrect = 1) item, and it asked participants to answer a single problem from a Raven's Progressive Matrices test of general intelligence (Figure 3). In order to solve the problem, participants needed to recognize that the lines are rotating, one small segment at a time, and choose one of 8 images that completed the pattern required to fill in the lower right block.

For this question error rates were more reasonable than for the other binary question (question 3), the question was challenging but not as close to impossible, with error rates mostly around .25. On round 3 the Proportion of Errors scores were: 6/25, 9/26, and 7/23 for the Basic Delphi, Enhanced Delphi, and Individual group judgment conditions respectively (non-significant with a standard Chi-squared test on the proportions (2 df) = 0.70, n.s. If we separate the results by location, there are no effects of judgment condition, although error rates are higher in the CRL, as might be expected: 3/14, 2/13, and 2/12 errors for Basic, Enhanced, and Individual, Chi-squared (2 df) = 0.19, n.s; and for the CRL data: 3/11, 7/13, and 5/11 errors, Chi-squared (2 df) = 1.76, n.s.

Analysis of Judgments for Question 8: What percentage of Americans has a pet? (answer: 62%)

First, recall that we included Questions 8, 9, and 10, because intuitively we felt that there was little room for group discussion and information pooling to improve performance. So, our hypothesis is that Enhanced Delphi will not show an advantage for these questions. Although, we might still expect that any Delphi method would increase accuracy slightly over the

Individual judgment condition. Second, these questions were added to the experimental protocol late in the study and so all the results are from the CRL.

There did seem to be systematic improvement across rounds for the Enhanced Delphi condition, but no others. No other patterns were obvious in the descriptive plots.

Figure 15. Trimmed box plot summary of the values for median accuracy for Q8 with outliers excluded.

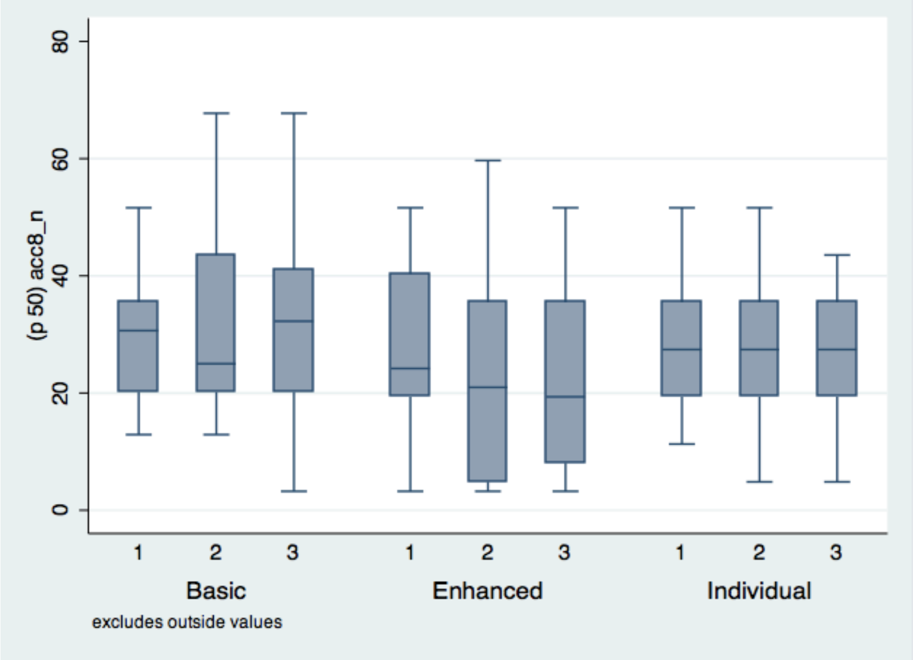
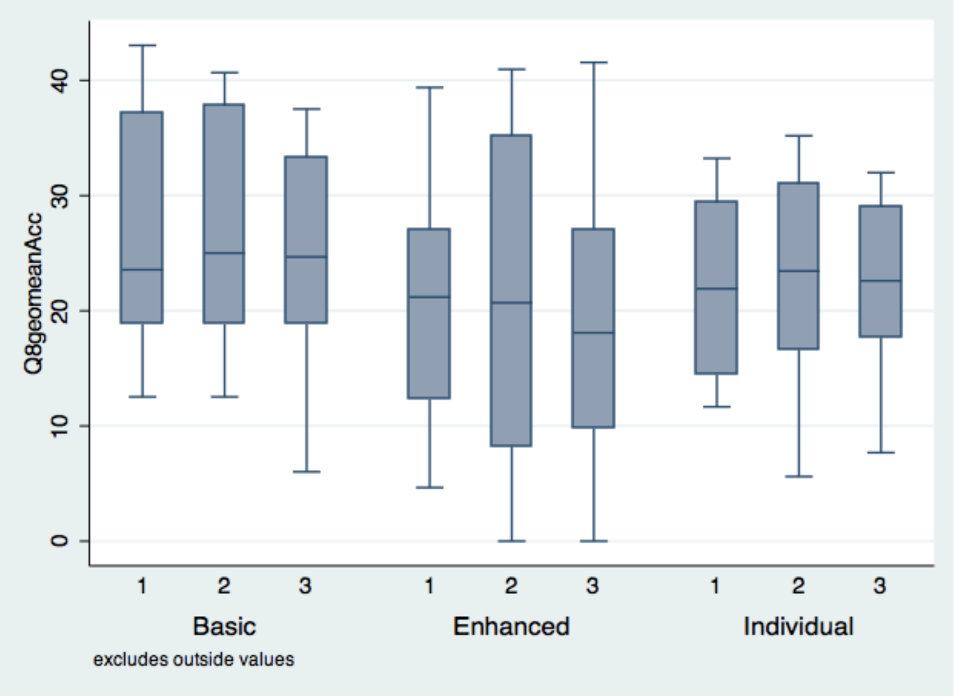


Figure 16. Trimmed box plot summary of the values for geometric mean accuracy for Q8 with outliers excluded.



The Kruskal-Wallis ranks test hinted that the Enhanced Delphi was most accurate (lowest rank-sum score); but the results were far from significant (K-W Chi-squared (2 df) = 1.83, n.s.). Of course, we need a larger sample, and more statistical power to reach a conclusion that there is a small or null effect on Question 8.

Table 14. Kruskal-Wallis one-way analysis of ordinal ranks only for round 3 estimates of question 8 (CRL and DRL).

Condition	Observations	Rank Sum
Basic	12	277.00
Enhanced	15	264.50
Individual	13	278.50

Analysis of Judgments for Question 9: How many murders were officially registered in the US in 2013? (answer: 14,827)

Here again, we are concerned that we may have elicited answers on different scales. For example, the apparent peak near the Percent Absolute Error score of 100, may indicate that participants were giving low numbers, with the intention that they be “scaled up” on a “thousands” scale. As with question 2, many answers were in the range less than 100, a pretty odd answer for “Murders in US in 2013” (Table 15). It might be argued that what participants were doing, again perhaps in good faith, was submitting answers in the thousands (e.g., 15 thousand would be a very good answer). The experimental procedure failed to constrain these responses, and so the answers to this question are potentially not interpretable.

Table 15. Ranges of Numerical Answers to Q9.

Range	Frequency	Percent	Cumulative %
0-10	20	5.10	5.10
11-100	59	15.05	20.15
< 100	143	79.85	100.00

In any case, there are no discernable patterns indicating between judgment condition differences in the data. Although there is the odd discrepancy in apparent variability across groups for the Basic Delphi groups on the two measures, Group Median Error Scores versus Group Geometric Mean Error Scores.

Figure 17. Trimmed box plot summary of the values for median accuracy for Q9 with outliers excluded.

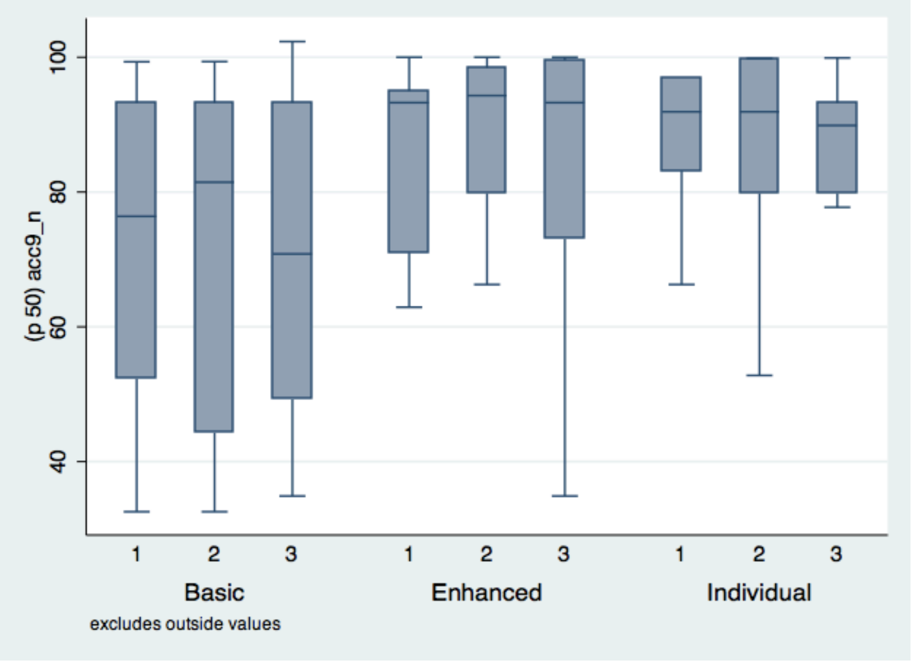
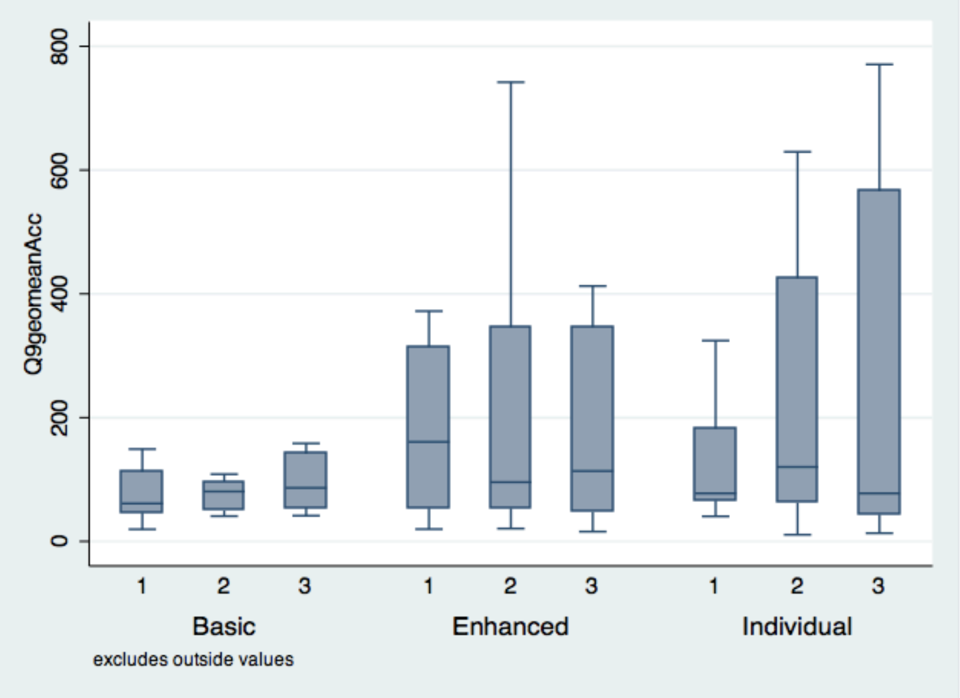


Figure 18. Trimmed box plot summary of the values for geometric mean accuracy for Q9 with outliers excluded.



The Kruskal-Wallis tests found no effects (although recall that this sample is small and composed mostly of CRL participants (K-W Chi-squared (2 df) = 1.94, n.s.). (We cannot compare results across locations as almost all the data is from the CRL.)

Table 16. Kruskal-Wallis ordinal ranks test only for round 3 estimates of question 9 (CRL and DRL).

Condition	Observations	Rank Sum
Basic	12	199.50
Enhanced	15	339.50
Individual	13	281.00

Analyses of Judgments for Question 10: In 1900, the percentage of the total US populations that was aged 65 and over was 4.1%. What was the percentage in 2013? (answer: 13.1%)

There are no patterns obvious in the descriptive plots.

Figure 19. Trimmed box plot summary of the values for median accuracy for Q10 with outliers excluded.

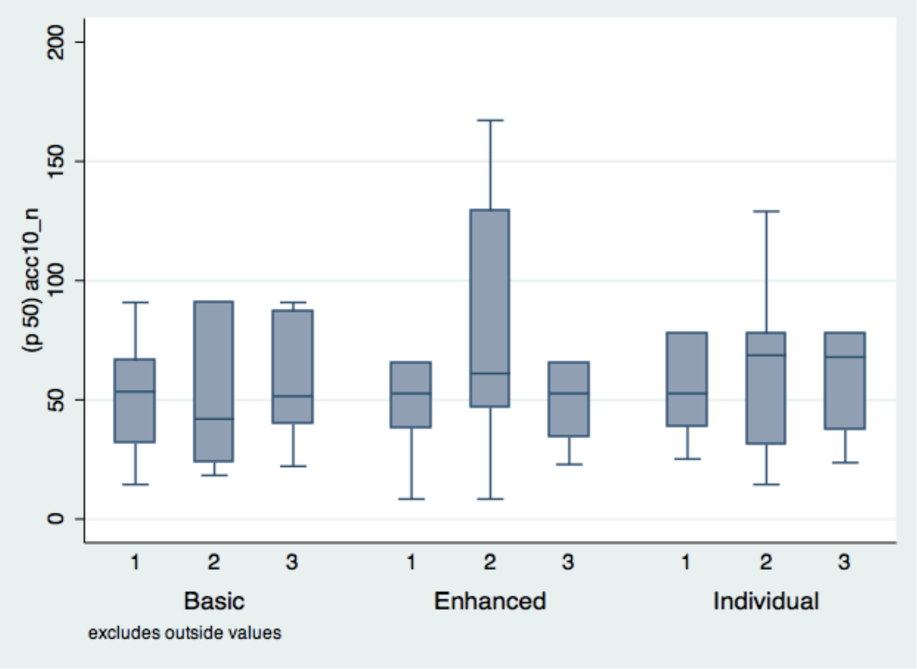
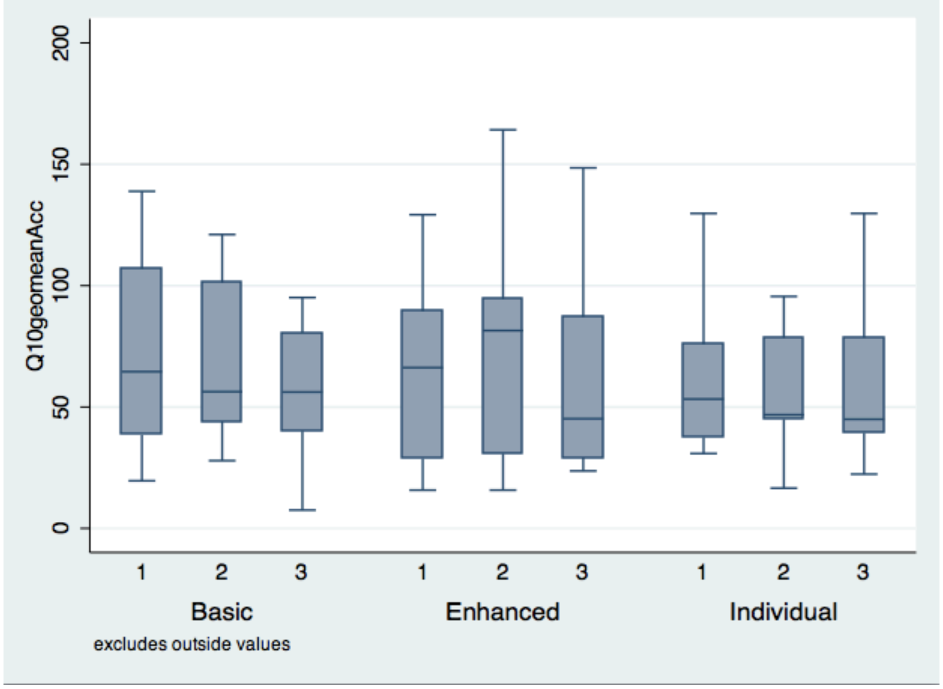


Figure 20. Trimmed box plot summary of the values for geometric mean accuracy for Q10 with outliers excluded.



The Kruskal-Wallis tests found no effects (although recall that this sample is small and composed mostly of CRL participants (K-W Chi-squared (2 df) = 0,65, n.s.)). (We cannot compare results across locations as almost all the data is from the CRL.)

Table 17. Kruskal-Wallis one-way analysis of ordinal ranks only for round 3 estimates of question 10 (CRL and DRL).

Condition	**Observations**	**Rank Sum**
Basic	12	254.00
Enhanced	15	279.50
Individual	13	286.50

Omnibus Chi-Squared Significance Test

If we consider the Study 2 results, from the Questions 1, 3, 4, 5, 6, 7[12] and comparing Enhanced versus Individual group judgment methods, we see that the performance of Enhanced groups on the median Percent Absolute Accuracy score (on Round 3) is always higher than the performance of Individual participants across all 6 of these questions. If we make the assumption that responses to the different questions are independent, we can combine the significance levels across the 6 questions. If we do so (following Fisher's combined probability test (1932, Winer, 1971, pp. 49-50), we calculate an omnibus Chi-squared significant test value of 21.50 (with 12 df), significant at the p < .05 level. We have made this calculation only for the sake of speculation. We do not believe that the independence assumption is met and, in current practice, more complex combination rules are usually prescribed (cf. Whitlock, 2005). And, we

[12] We drop Question 2 because of the methodological problem of absurdly low numerical responses that probably resulted from a misunderstanding of the response scale (participants responded on a "millions scale," but the question was scored on a "1s scale"). We do not include Questions 8, 9, and 10 because they were included, as a control, on the expectation that they would mute or eliminate the advantage of enhanced information pooling.

do not believe that the current results demonstrate the reliability of the Enhanced versus Individual conditions difference at a level of confidence implied by a $p < .05$ significance level. As we have said elsewhere, additional data needs to be collected to draw convincing conclusions about the reliability of any differences in performance observed between our group judgment conditions.

Description of the Types of Information Being Shared in the Enhanced Delphi Method

In the Introduction and Training forms we provided model answers and insights for two questions, one that asked how many ping-pong balls can fit inside a Boeing 747 and a second that asked how many piano tuners there are in greater Chicago. We also encouraged participants to share, "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?" Our review of participants' shared insights (Table 18) suggests some participants followed such guidance while others shared more mundane rationales for why they gave this or that answer.

Table 18. Examples of Shared Insights

#	Example Insight
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 [sic] 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 [sic] 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 [sic] mentioned after nearly every mass shooting and that half are suicides half homicides."
9	"Theres [sic] about 300 to 400 murders in big Metropolitan areas like Chicago and New york per year. then [sic] add in Detroit and that might be another 500 or so and then a bunch in LA. then [sic] a bunch in the rest of the country."
10	"Our population has nearly tripled in size, so I multiplied [sic] 4.1 by 3 to get 12.3."

The next table (Table 19) provides some examples of the back-and-forth “dialog” about one of the questions asked in study 2 for one group answering question 4 about how many Chicago ‘L’ stations there are.

Table 19. Dialogue Among Members of One Enhanced Delphi Team about Question 4

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 participant 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.

Study 2: Discussion of Results

Study 2 can be interpreted, with caution, as supporting the hypothesis that information pooling increases group judgment accuracy. Table 20 is an overview, question-by-question summary of the findings from Study 2, and if any group process method is the most accurate, it is Delphi-Enhanced, where information-pooling was supported most fully by the judgment procedure. Delphi-Enhanced seems to work both because outliers are somewhat reduced and because, when we only consider non-outlier data, there is a shift towards increased accuracy. Furthermore, Delphi-Enhanced requires that the individual participants be engaged in the task and somewhat knowledgeable. The Delphi-Enhanced advantage is driven mostly by data from the more elite University of Chicago participants.

Table 20. Summary table of results.

#	Question	Condition Effects on Round 3 Accuracy	Comment re: DRL versus CRL Comparisons	Significant Results	Other Observations
1	Height of tallest man-made building (answer: 2722 ft)	Delphi-Enhanced is best	larger number of outliers in CRL (15%) versus DRL (1%)		
2	Contiguous area of US (3,114,00 sq. mi.)	Delphi-Basic and -Enhanced (tied), followed by Individual	no differences		possible that participants were answering on a <u>millions</u> scale so question may not be interpretable
3	Starbucks in US (10,924)	Delphi-Basic is best	same pattern across locations		
4	Chicago 'L' stations (145)	Delphi-Basic is best	Delphi-Enhanced better in DRL	DRL: K-W Chi-squared (2 df) = 4.75, p < .10	
5	Metro with most foreign-born people (Miami)	Delphi-Enhanced and Individual (tied), followed by Delphi-Basic	Delphi-Enhanced better in DRL	DRL: K-W Chi-squared (2 df) = 6.50, p < .05	binary answer; difficult question (error rates over 80% in all conditions)

Table 20, continued

6	Driving time from WY to FL with stops (83 hr.)	Delphi-Enhanced	same pattern across locations		
7	Raven's IQ	Delphi-Basic	higher error rates in CRL		binary answer; difficult but not impossible question (error rates around .25)
8	Americans who own a pet (62%)	Delphi-Enhanced	almost all data from CRL		
9	Murders in US (14,827)	no discernable ordering	almost all data from CRL		possible that some participants were answering in thousands so question may not be interpretable
10	US Population aged 65 and over (13.1%)	no discernable ordering	almost all data from CRL		

However, this conclusion must be qualified by a great deal of uncertainty. First, the conclusion is not perfectly universal across the 7 Questions that were selected to capitalize on information pooling. Second, the samples are still too small to yield many reliable conclusions. Third the data is complicated and includes many informative outliers. Third, it looks like a major ingredient in the recipe that produces the information-pooling advantage is highly motivated, somewhat knowledgeable, group members.

The bottom line conclusion on Study 2 is to encourage future research, with larger samples and a couple of small methodological refinements. We conclude that there is strong suggestive evidence for the value of information pooling, the Delphi-Enhanced method, in making certain classes of judgments.

Why the outliers are more than just a methodological inconvenience

One of the primary mechanisms that produces "wisdom of crowds" accuracy effects is error damping. In real applications of such methods, a very common obstacle to efficient performance is the occasional poorly-informed or unlucky "crazy solution." Therefore, outliers are a very real practical concern in designing a collective intelligence method. And, any social process that reduces error-prone outliers has practical value. Of course, we also hoped to find evidence for a more conceptually interesting process advantage for information-pooling mechanisms. And, there is some suggestive evidence for that mechanism in addition to simple error-damping.

Methodological lessons from Study 2

One important lesson was learned from the substantial numbers of small numerical responses for Questions 2 and 9. These almost certainly indicate a methodological failure. The participants, who were registering absurdly low estimates, were surely responding on a different "subjective metric" (in the millions or the thousands) than the one requested by the problem instructions. The remedy is obvious, in hindsight, and future research must take even more care to get responses from all participants on the same metric scales.

The second methodological lesson has more of a substantive content. Highly motivated, relatively well-informed participants are necessary to demonstrate the values of an intellectually-demanding judgment process, like one that depends on effective information-pooling. We label this lesson more substantive, because it encourages constructive thought about how to select judges to make important estimates, and how to incentivize vigorous participation.

Conclusion

Our current research program might be conceptualized as a quest for the illusive information pooling advantage. In this current piece of work, we used a Delphi-Enhanced Method to promote systematic pooling of judgment-relevant information. As noted above, we believe Study 2 can be interpreted, with caution, as supporting the hypothesis that information pooling increases group judgment accuracy. We would like to conclude by addressing directly what we believe is the "recipe" for information pooling advantage, and more specifically how we might further explore the validity of that recipe. Relatedly we would then like to note how, practically speaking, one might determine where and when information pooling will work. Then we will note scope conditions—qualifications on our conclusions—before wrapping up with a discussion of future avenues of research.

The Recipe for the Information Pooling Advantage

We conclude that our results, especially those relating to Study 2, provide encouraging—but as yet tentative—information about "where" to find the advantage of information pooling: (i) Judgment items with unshared relevant information; (ii) incentives to pool information; and

(iii) knowledgeable participants. Here is a bit more on these sources of advantage in terms of information pooling with a focus on how we might further explore the importance of each factor:

In terms of judgment items with unshared relevant information, perhaps an obvious “next step” would be to utilize hidden profile questions in a Delphi Method-based experimental procedure. Such an approach would bring a degree of formality to the issue of how much information is shared versus unshared for a given question, and would have the potential to shed light on the importance of shared versus unshared information as they relate to information pooling.

When it comes to incentives to pool information, we can relatively easily move beyond the simple—perhaps crude—incentives we offered in Study 2, where individual participants were offered prizes if they were on the most accurate team or were the most accurate individual. With some effort, it would be possible to offer real-time rewards based on the quality of shared information provided to one’s teammates. This could be accomplished via matching, potentially in real-time, submitted insights to the kinds of insights known to be correlated with improved team performance, perhaps from prior research or from expert experience.

Knowledgeable participants are one key to the success of information pooling. We were surprised at the difference between the responses from the two pools of participants we utilized in Study 2, those recruited by the downtown Chicago research laboratory (CRL) versus the Hyde Park research laboratory (DRL). What needs to be determined is whether information pooling can be made to work among participant pools like the CRL, or whether the advantages of information pooling can brought to the fore with elite, highly motivated participants like those at the DRL.

What next?

The obvious next step in our research agenda is to continue Study 2 but in a refined form. That is to say, keep running Study 2 but with more subjects, a bigger emphasis on highly motivated (DRL-type subjects), and other methods improvements. Methodological improvements would include addressing the fact that on questions like 2 and 9, different subjects appear to be answering the question on different scales than was intended (millions and thousands respectively for questions 2 and 9). Any subsequent studies we conduct will ensure all participants are using the same metric scales. One possible solution to this particular challenge is to require that such answers submitted via a Google Form fall within a certain range. With modifications such as this in place, and with additional subjects—especially like those from the DRL—we believe we will see statistically significant results in support of Enhanced Delphi for a majority of questions 1 through 7, and will be able to convincingly demonstrate that information pooling does not work well for judgments like those requested in questions 8 through 10.

We would also like to futher explore the specific content of shared information. We see value in completing a descriptive analysis that studies the characteristics of insights and rationales that are more (less) effective in terms of improving group performance or that are associated with individuals who make more accurate judgments than others. Such analysis might be either quantitative or qualitative (or both).

Simple quantitative analysis might involve counting number of words or characters in a given insight. More sophisticated, but still relatively simple qualitative analysis of participant-provided insights could be completed using a text analysis software application. One well-regarded application is called Linguistic Inquiry and Word Count (LIWC). The software was designed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis and has been used by other researchers to shed light on contents like the shared rationales in our experiment. LIWC is able to calculate the degree to which people use different categories of words across a wide array of texts. The software can assess 70 or so dimensions including self-references, big words, and positive versus negative affect, all of which would be informative if tied to higher performance in terms of individual or group accuracy.

We also see value in analyzing the manner in which content is shared and how it affects information pooling and ultimately accuracy. A future study might shed light on the relative value of sharing insights, as well as numeric data, using simple written characters and sentences versus sharing visualizations such as box plots, histograms, pie charts, etc.

Finally in terms of where to go next, we suggest extending the Delphi-Enhanced Method to make it even more effective. We provided participants our suggestions for what a good insight looks like, and we further encouraged the production of quality insights via two training questions with example insights. Still, we can certainly go to greater lengths to improve the quality of the Delphi-Enhanced Method. One possibility is to encourage participants to share multiple different kinds of insights, e.g., insights based on functional expertise (“I’m an architect”), insights based on local knowledge (“I was just in Dubai last week and took a trip to the observation deck of the Burj Khalifa”), insights based on cognitive ability (“Based on two team-mates thinking the tallest building had 162 floors, I felt the height must be 162 x 15 = 2,430 feet”) or aptitude for reacting to and intelligently integrating previous round answers (“I noticed that one team-mate appeared especially knowledgeable about tall buildings”).

We believe the next generation of collective intelligence methods will be based on improved information pooling methods, and will focus on enhancing Stage 1 information acquisition. We believe the present research makes a solid contribution to this evolution of "designed collective intelligence methods" and, with other scholars in this field, we look forward to expanding the boundaries of the scientific understanding of Stage 1 information acquisition.

References

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. Technological Forecasting and Social Change, 78(9), 1500–1513. doi:10.1016/j.techfore.2011.07.007

Bolger, F., Stranieri, A., Wright, G., & Yearwood, J. (2011). Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? Technological Forecasting and Social Change, 78(9), 1671–1680. doi:10.1016/j.techfore.2011.06.002

Clemen, R.T., & Winkler, R.L. (1999). Combining probability forecasts from experts in risk analysis. Risk Analysis, 19(2), 187-203.

Dalkey, N., Helmer, O., & CA, R. C. S. M. (1962). An Experimental Application of the Delphi Method to the Use of Experts.

Davis-Stober, C.P., Budescu, D.V., Dana, J., & Broomell, S.B. (2014). When is a crowd wise? Decision, 1(2), 79-101.

Davis, J.H., Laughlin, P.R. & Komorita, S. (1979) The social psychology of small groups: Cooperative and mixed-motive interaction. Annual Review of Psychology, 27, 501-541.

Dawid, A.P., DeGroot, M.H., & Mortera, J. (1995). Coherent combination of experts' opinions. TEST, 4(2), 263-313.

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences*, *108*(36), E625–E625. doi:10.1073/pnas.1109947108

Felicity, H., & Sinead, K. (2011). Enhancing rigour in the Delphi technique research. Technological Forecasting and Social Change, 78(9), 1695–1704. doi:10.1016/j.techfore.2011.04.005

Fisher, R.A. (1932). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.

Fraidin, S. N. (2004). When is one head better than two? Interdependent information in group decision making. *Organizational Behavior and Human Decision Processes*, *93*(2), 102–113. doi:10.1016/j.obhdp.2003.12.003

Galton, F. (1907). Vox populi. *Nature, 75,* 7-9.

Galton, F. (1907). The ballot-box. *Nature, 75,* 509-510.

Gigone, D., & Hastie, R. (1997a). Proper analysis of the accuracy of group judgments. Psychological Bulletin, 121(1), 149. doi:10.1037/0033-2909.121.1.149

Gigone, D., & Hastie, R. (1997b). The impact of information on small group choice. Journal of Personality and Social Psychology, 72(1), 132. doi:10.1037/0022-3514.72.1.132

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

Green, K. C., Armstrong, J. S., & Graefe, A. (2007). Methods to elicit forecasts from groups: Delphi and prediction markets compared. Foresight: the International ….

Gürçay, B., Mellers, B. A., & Baron, J. (2014) The power of social influence on estimation accuracy (submitted for publication).

Hackman, J. R, & Morris, C.G. (1975) Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. Advances in Experimental Social Psychology, 8, 45-99.

Hackman, J. R, Jones, L. E., & Mcgrath, J.E. (1967) A Set of Dimensions for Describing the General Properties of Group-generated Written Passages. Psychological Bulletin, vol 67(6), 379-390.

Hackman, J. R. & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. Organizational Behavior and Human Performance, 16, 250-279.

Hackman, R.J. (1968) Effects of Task Characteristics on Group Products, Journal of Experimental Social Psychology, 4, pp. 162-187.

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.

Hastie, R., & Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. Psychological Review, 112(2), 494. doi:10.1037/0033-295X.112.2.494

Henry, R. A. (1995). Improving group judgment accuracy: Information sharing and determining the best member. Organizational Behavior and Human Decision ….

Hogarth, R.M. (1977). Methods for aggregating opinions. In H. Jungermann & G. DeZeeuw (Eds.), Decision Making and Change in Human Affairs (pp. 231-255). Dordrecht, Netherlands: Reidel.

Jacobs, R.A. (1995). Methods for combining experts' probability assessments. Neural Computation, 7(5), 867-888.

Janis, I. L. (1972). Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.

Landeta, J. (2006). Current validity of the Delphi method in social sciences. Technological Forecasting and Social Change, 73(5), 467–482. doi:10.1016/j.techfore.2005.09.002

Laughlin, P. R. (1980) Social combination processes of cooperative problem-solving groups on verbal intellective tasks. In M. Fishbein (Ed.), Progress in social psychology (Vol. 1). Hillsdale, N.J.: Erlbaum.

Linstone, H. A., & Turoff, M. (Eds.). (1975). Delphi Method: Techniques and Applications. Addison-Wesley Educational Publishers Inc.

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

McGrath, J. E. (1984). Groups: Interaction and Performance. Inglewood, N. J.: Prentice Hall, Inc.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gürçay, B., Fincher, K., et al. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. Psychological Science doi:10.1177/0956797614524255

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

Rowe, G., & Wright, G. (2011). The Delphi technique: Past, present, and future prospects â€” Introduction to the special issue. Technological Forecasting and Social Change, 78(9), 1487–1490. doi:10.1016/j.techfore.2011.09.002

Rowe, G., Wright, G., & McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. Technological Forecasting and Social ….

Shaw, M. E. (1976). Group dynamics: the psychology of small group behavior. 2d ed. New York: McGraw-Hill.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. Journal of Personality and Social Psychology, 48(6), 1467. doi:10.1037/0022-3514.48.6.1467

Steiner, I. D. (1966) Models for inferring relationships between group size and potential group productivity. Behavioral Science, 11, 273-283.

Steiner, I.D. (1972) Group processes and productivity. New York: Academic Press.

Van de Ven, A. H., & Delbecq, A. L. (1974). The effectiveness of nominal, Delphi, and interacting group decision making processes. *Academy of Management Journal*.

Wentholt, M., Fischer, A., Rowe, G., & Marvin, H. (2010). ScienceDirect.com - Food Control - Effective identification and management of emerging food risks: Results of an international Delphi survey. Food Control.

Whitlock, M.C. (2005). Combining probability from independent tests: The weighted z-method is superior to Fisher's approach. Journal of Evolutionary Biology, 18(11), 1368-1373.

Winer, B.J. (1971, 2nd Edition). Statistical Principles in Experimental Design. New York: McGraw-Hill.

Woudenberg, F. (1991). An evaluation of Delphi. Technological Forecasting and Social Change, 40(2), 131–150.

Yaniv, I. (2004). ScienceDirect.com - Organizational Behavior and Human Decision Processes - Receiving other people's advice: Influence and benefit. Organizational Behavior and Human Decision ….

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. Organizational Behavior and Human Decision Processes, 83(2), 260–281. doi:10.1006/obhd.2000.2909