ITAS - EPP
Elektronische Pre-Prints

Andreas Graefe / J. Scott Armstrong

Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task

Pre-Print: 17.03.2010

Erschienen in: International Journal of Forecasting 27(2011)1, S. 183-195

ITAS - Elektronische Pre-Prints

Allgemeine Hinweise

Wie mittlerweile viele wissenschaftliche Einrichtungen, bietet auch ITAS elektronische Pre-Prints an, die bereits zur Publikation akzeptierte wissenschaftliche Arbeiten von Mitarbeiterinnen und Mitarbeitern - in der Regel Buchbeiträge - darstellen.

Für die Autoren bietet dies den Vorteil einer früheren und besseren Sichtbarkeit ihrer Arbeiten; für die Herausgeber und Verlage die Möglichkeit einer zusätzlichen, werbewirksamen Bekanntmachung des jeweiligen Buchprojekts. Auf die in Aussicht stehende Veröffentlichung wird hingewiesen. Nach Erscheinen der Publikation werden der geänderte Status vermerkt und die bibliographischen Angaben vervollständigt.

Allgemeine Anregungen und Kommentare zu den ITAS Pre-Prints richten Sie bitte an Fritz Gloede (fritz.gloede@kit.edu).

Empfohlene Zitierweise des vorliegenden Pre-Prints:

Graefe, A.; Armstrong, J. S.: Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task.
Karlsruhe: ITAS Pre-Print: 17.03.2010;
http://www.itas.fzk.de/deu/lit/epp/2010/grar10-pre01.pdf

Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task

Accepted for publication in the *International Journal of Forecasting*

Andreas Graefe
Institute for Technology Assessment and Systems Analysis
Karlsruhe Institute of Technology, Germany
graefe@kit.edu

J. Scott Armstrong
The Wharton School
University of Pennsylvania, Philadelphia, PA
armstrong@wharton.upenn.edu

Feb 1, 2010

Abstract. We conducted laboratory experiments to analyze the accuracy of three structured approaches (nominal groups, Delphi, and prediction markets) compared to traditional face-to-face meetings (FTF). We recruited 227 participants (11 groups per method) that had to solve a quantitative judgment task that did not involve distributed knowledge. This task consisted of ten factual questions, which required percentage estimates. While, overall, we did not find statistically significant differences in accuracy between the four methods, the results differed somewhat at the individual question level. Delphi was as accurate as FTF for eight questions and outperformed FTF for two questions. By comparison, prediction markets were unable to outperform FTF for any of the ten questions but were inferior for three questions. The relative performance of nominal groups and FTF was mixed and differences were small. We also compared the results from the three structured approaches to prior individual estimates and staticized groups. The three structured approaches were more accurate than participants' prior individual estimates. Delphi was also more accurate than staticized groups. Nominal groups and prediction markets provided little additional value compared to a simple average of forecast. In addition, we examined participants' perceptions of the group and the group process. Participants rated personal communication more favorable than computer-mediated interaction. Group interaction in FTF and nominal groups was perceived as highly cooperative and effective. Prediction markets were rated least favorable. Prediction market participants were least satisfied with the group process and perceived their method as most difficult.

Keywords: forecast accuracy, group decision-making, method comparison, satisfaction, combining

In situations where a lack of appropriate or available information precludes one from using quantitative methods, it can be helpful to incorporate human judgment to improve the forecast. But how does one get the best forecast when aggregating information from a group of people? While organizations most commonly rely on unstructured face-to-face meetings, it is difficult to find evidence to support the use of this strategy (Armstrong 2006). The literature suggests that structured approaches like nominal groups or Delphi provide for more accurate forecasts than traditional meetings.

In recent years, prediction markets gained interest as an approach to elicit information from people and a number of organizations have started to experiment with them. However, to date, we do not know much about the performance of prediction markets. Available studies are limited and often of a small scale. In particular, we do not know of any study that analyzed their relative performance compared to meetings as well as to other structured group techniques for aggregating the knowledge in groups. Since the emergence of the field, there is no meta-analysis that analyzed prediction markets' accuracy.

To address this deficit, we conducted laboratory experiments comparing unstructured meetings, nominal groups, Delphi, and prediction markets. We analyzed the relative accuracy of the four group techniques on a quantitative judgment task and examined participants' perceptions of their group and their group process.

Group judgment tasks

We selected problems that did not involve highly distributed information. This is not to suggest that problems involving highly dispersed information among participants are unimportant, but merely that we addressed a limited – but typical – forecasting situation: general problems for which people are unlikely to possess specific knowledge. In our study, groups had to come up with estimates for a quantitative judgment task. This task consisted of ten factual questions that required percentage estimates. These questions had correct solutions but all group members could be expected to have some uncertainty about the answers.

Findings from group performance studies are generally task-specific. The critical question arises whether the tasks used in experimental settings are representative of 'real-world' problems and thus allow for generalizations to different types of tasks (like forecasting problems) or participants. For example, Wright and Ayton (1986) provided some evidence that the findings from calibration studies that used general knowledge questions are not applicable to judgmental forecasting problems, since both task types involve different levels of uncertainty. Furthermore, it is often argued that student participants in laboratory experiments are laymen, whereas participants in 'real-world' forecasting scenarios are experts who are expected to possess superior knowledge.

In general, for a judgmental forecasting method to perform well, it should be able to aggregate efficiently the relevant information of its members. Thus, a basic precondition for the generalizability of

problems used in experimental settings is the involvement of participants in trying to solve them. Factual questions are commonly analyzed by group decision-making as they have been shown to spark the interest of student participants.

Group techniques

We analyzed four group techniques that differ in the amount and structure of interaction permitted between group members: face-to-face meetings, nominal groups, the Delphi method, and prediction markets.

Unstructured face-to-face meetings

Unstructured face-to-face meetings (FTF) allow any form of direct interaction between group members. Although meetings are most common for group decision-making in organizations, they have been shown to be subject to many biases and drawbacks. For example, (1) it requires time and effort for a group to maintain itself; (2) groups tend to aim at reaching 'speedy decisions', do not consider all problem dimensions, and thus tend to pursue a limited train of thought, which leads to a 'central tendency effect' or 'groupthink'; (3) less confident group members, or people from lower hierarchy levels, may keep silent about their reservations because of group pressures for conformity or implied threats of sanctions; (4) dominant personalities tend to exert excessive influence on the group. For a summary of these issues see Van de Ven and Delbecq (1971).

In sum, Armstrong (2006) found little evidence to support the use of meetings for forecasting or decision-making. In addition, meetings are expensive to schedule and to run. In some situations, however, one must be concerned not only with the quality of a decision but also with its acceptability. People generally enjoy human interaction and the sense of working together and meetings have been shown to often achieve high levels of satisfaction (e.g. Van de Ven & Delbecq 1974, Boje & Murnighan 1982).

Nominal group technique

The *nominal group technique (NGT)*, developed by Van de Ven and Delbecq (1971, 1974), tries to account for some of the drawbacks of traditional meetings by adding a structured format to direct interaction. This process is conducted in three steps: First, group members work independently and generate individual estimates on a problem. Then, the group enters unstructured discussion to deliberate on the problem. Finally, group members work again independently and provide their final individual estimates. The group result is the aggregated outcome of these final individual estimates. The literature refers to this process also as *estimate-talk-estimate* (Gustafson et al. 1973).

The idea of NGT is that direct interaction during the assessment or evaluation phase can have a positive impact on estimation or problem solving. In particular, it can help group members to clarify and

justify their points of view, which may help the group to make more informed decisions. Nonetheless, in the phases of generating answers and making the final decisions, NGT prevents direct interaction between group members to reduce the drawbacks known with traditional meetings. See Van de Ven and Delbecq (1971) for a summary of advantages of nominal groups.

The Delphi method

Delphi is a multiple-round survey in which participants anonymously reveal their individual estimates as well as comments on a problem. After each round, the individual estimates and comments are summarized and reported as feedback to participants. Taking into account this information, participants provide their new estimate in the following round. The group result is the aggregated outcome of the individual estimates of the final round. Unlike FTF or NGT, which require physical proximity of group members, participants in *Delphi* are physically dispersed and do not meet in person. In general, Delphi functions somewhat similar to NGT. The main difference is that written interaction is utilized during the whole process to avoid direct interaction between group members.

The strengths of the method are seen in its structured communication process that enables discussion and helps a group to achieve consensus but limits the drawbacks associated with direct interaction. For reviews of Delphi see Woudenberg (1991) and Rowe and Wright (1999).

Prediction markets

Prediction markets were popular in the late 1800s (Rhode & Strumpf 2004) and have an impressive track record in forecasting elections. In analyzing 964 polls for the five presidential elections from 1988 to 2004, Berg et al. (2008) found that the respective forecasts of the *Iowa Electronic Markets* were closer to the actual election results than individual polls 74% of the time. However, this advantage disappeared when polls were averaged and damped (Erikson & Wlezien 2008).

Prediction markets are gaining attention in various fields of forecasting. The idea is to set up a contract whose payoff depends on the outcome of an uncertain future event. This contract, which can be interpreted as a bet on the outcome of the underlying future event, can then be traded by participants. As soon as the outcome is known, participants are paid off in exchange for the contracts they hold. Based on their individual performance, participants can win money. If one thinks the current group estimate is too low (high), one will buy (sell) stocks. Thus, through the prospect of gaining money, participants have an incentive to become active in the group process whenever they expect the group estimate to be inaccurate. As in Delphi, participants are mutually anonymous and thus not subject to social pressures afflicted with direct interaction. The main difference is that participants exchange information continuously through the price signal of the market but do not share comments or reasons for why they buy or sell a contract. For an

explanation of the concept of prediction markets and a useful summary of the method see Wolfers and Zitzewitz (2004).

Prediction markets using play-money have also proven to be accurate. In comparing the results of play-money and real-money markets for sport events, Servan-Schreiber et al. (2004) did not find differences in accuracy. Yet, Rosenbloom and Notz (2006) found real-money markets to be more accurate for non-sports events.

Related work

Structured group techniques are expected to limit biases associated with meetings and prior research seems to support this assumption. In their review, Rowe and Wright (1999) reported superior accuracy of Delphi over unstructured interaction by a score of five studies to one. In excluding two papers from their analysis that did not analyze accuracy for quantitative judgment tasks, we adjusted the score to three studies to one, still with two ties. Similarly, in summarizing the literature, Woudenberg (1991) found an – albeit slight – superiority of Delphi over unstructured interaction. Furthermore, again for quantitative judgment tasks, Gustafson et al. (1973) found NGT to be more accurate than FTF, although Fischer (1981) found no differences. Evidence on the relative performance of NGT, Delphi, and prediction markets for quantitative judgment tasks is scarce. While Gustafson et al. (1973) showed that NGT was more accurate than Delphi, two studies found no differences (Fischer 1981, Boje & Murnighan 1982). In the case of prediction markets, we do not know of any work that compared the method to FTF, NGT or Delphi.

Little work has been done on analyzing participants' perceptions of group processes, although this can be important for the acceptance of its results. If group members or decision-makers feel dissatisfied with the process, its outcome may not be adopted – even if highly accurate. Vice versa, a process that is highly satisfying for participants may not necessarily lead to accurate results. In general, personal interaction in groups can either lead to coherence, and thus high perceived satisfaction, or disagreement, resulting in frustrated group members. For quantitative judgment tasks, Boje and Murnighan (1982) analyzed participants' perceptions for Delphi, NGT, and people working alone. They found that NGT was rated most favorable in terms of effectiveness, satisfaction, and freedom to participate, whereas Delphi was rated only slightly superior compared to working alone. Van de Ven and Delbecq (1974) compared participants' perceptions of nominal groups, Delphi, and unstructured meetings for an idea generation problem. They found that NGT participants expressed highest satisfaction with the process whereas differences between Delphi and meetings were small.

For prediction markets, we do not know of any study that analyzed how people perceive participation and how this might affect acceptability of market results. However, practical experience indicates that people have problems in understanding how prediction markets work. In particular, they have problems in translating information into market prices (Green et al. 2007). In addition, people appear

to have difficulties in understanding what the prices mean For example, during the 2008 financial crisis, intrade.com launched a contract to predict whether the U.S. government will pass a bail out. At one time, the contract was traded at a price of $80, which means that the market forecasted an 80% probability that the bail out will go through. A staff writer at CNNMoney.com – who could be expected to possess a certain understanding of financial instruments – wrongly interpreted the market prices in stating that *80% of the participants thought* that the event will come true (CNNMoney.com: *Oddsmakers see bailout OK soon*, September 24, 2008). Misunderstanding of how prediction markets work might be critical for their acceptance by participants as well as decision-makers. If people do not understand how to reveal information, how information is aggregated into market prices, and how to interpret market prices, it is doubtful that they have favorable perceptions of the method and, thus, have confidence in its results.

Hypotheses

Given the lack of evidence, we had no directional hypothesis on the relative accuracy of the three structured approaches (NGT, Delphi, and prediction markets) since arguments could be found in either direction. For example, prediction markets might be advantageous to Delphi because traders can continuously update their estimates and are not tied to a limited number of rounds. On the other hand, market participants do not reveal reasons for their estimates. Thus, other group members might not be able to relate to *why* the group estimate has changed. Accordingly, the possibility of argumentation and reasoning might favor NGT and Delphi. For accuracy, we expected: NGT ≈ Delphi ≈ prediction markets > FTF.

We sent the study design to experts in the field and asked them to provide their prior estimates on the relative accuracy of the methods. To assure that the experts were familiar with all four methods, we primarily reached out to people experienced with prediction markets. Eight experts[1] responded and we derived the following ranking from their priors: Delphi > prediction markets > NGT > FTF. Expecting that all three structured approaches would outperform FTF, the expert priors were consistent with our hypotheses. In particular, none of the experts expected FTF to do better than any of the structured approaches.

To obtain a measure for the acceptability of the methods, we asked for participants' perceptions of the group and the group process along seven attributes. In particular, participants rated (1) cooperation and (2) disagreement among group members and revealed their (3) confidence in the group results. They also provided ratings of how (4) free they felt to participate, whether their (5) time was well spent, how (5) difficult it was to participate, how (6) satisfied they were with the group process and how (7) effective they thought the group process was for solving the estimation problem. Since our group interaction

[1] Thanks to Yiling Chen, Kesten Green, Robin Hanson, Stefan Luckner, Gene Rowe, Martin Spann, Gerrit Van Bruggen and Eric Zitzewitz.

processes simulated non-hierarchical meetings, we expected group pressures to be less. Assuming that people generally enjoy human interaction and the sense of working together, we hypothesized that personal interaction in FTF and NGT will lead to more favorable ratings compared to Delphi and prediction markets. Furthermore, due to less conformity pressures in NGT, we expected NGT to be rated more favorable than FTF; a result that had been confirmed earlier by Van de Ven and Delbecq (1974). In addition, since we expected the majority of traders to be unfamiliar with the process of revealing one's opinion by trading stocks, we expected prediction markets to be rated least favorable, in particular in terms of difficulty. In sum, for acceptability, we expected: NGT > FTF > Delphi > prediction markets.

Methods

Participants

Students (n=227) at the University of Pennsylvania, recruited by the Wharton Behavioral Lab, were randomly assigned to 44 heterogeneous groups – 11 per method. The 11 prediction market groups were further divided into two treatments. In 5 groups, traders immediately started trading. In 6 groups, traders independently generated their individual estimates before participating in the market.

Group size was determined by the number of students showing up in each session. Most groups (42 of 44) consisted of 4 to 6 subjects. One NGT group contained 3 subjects, one prediction markets group consisted of 7 subjects. Appendix 1 provides an overview of group size and (average) number of participants per method. All appendices can be found in a supporting online document: http://tinyurl.com/method-comparison-appendix.

All sessions lasted for one hour. Each participant was paid a $10 show-up fee. In addition, participants were remunerated depending on their group or individual performance. On average, $15 was distributed among each group. In FTF, NGT, and Delphi, $50 were equally distributed among participants of the two most accurate groups, the next two most accurate groups received $25, and the next $15. To meet the traditional pay-off mechanism of prediction markets, traders were paid based on individual performance, i.e. $6 for the best performing trader in each group, $5 for the second best, and $4 for the third best.

Materials and procedures

Data were collected in the Wharton Behavioral Lab. FTF and NGT were conducted in a small meeting room. Delphi and prediction market groups worked in a computer room on individual workstations, divided by partition walls to prevent personal communication. Each participant received general instructions explaining the relevant group technique as well as the respective pay-off mechanism. When the process started, forms were handed out to each participant for making personal notes and analyses. During the whole process, whenever individuals or groups were asked to reveal estimates, they had to

state their confidence in these estimates on a seven point numerical scale (1: not at all confident; 7: extremely confident). After the group interaction, participants of all group techniques received a post-task form to reveal their final individual estimates. In addition, the post-task form was used to obtain participants' perceptions of the group and the group process as a whole, again on a seven point numerical scale. Examples of the used materials can be found in Appendices 3 to 6.

Unstructured face-to-face meetings

Group members were seated around a table to discuss the problem with the goal of reaching a group estimate. Each group received one group questionnaire for their final group estimates. In addition, each group member received an individual questionnaire to note the achieved group estimates.

Nominal groups

First, group members were spread over three tables to work independently and to provide individual estimates for each of the ten questions on a paper questionnaire. After each participant had completed the questionnaire, group members were seated around one table to discuss their individual estimates with the group. Each group received one group questionnaire for their final group estimates. In addition, each group member received an individual questionnaire to note the achieved group estimates. Finally, group members were again spread out over three tables to provide their final individual estimates on the post-task questionnaire. The group results were calculated as the median of the final individual estimates.

Delphi

Before logging into the system, participants watched a video tutorial (http://tinyurl.com/method-comparison-Delphi) of how to use the Delphi software. To conduct our research in a realistic environment that can be adopted by practitioners, we used the free Delphi software available at ForPrin.com. This software was developed to educate users about Delphi and to aid them in the use of the standard Delphi procedure by including its four key features: anonymity of participants, iteration, controlled feedback, and statistical aggregation of individual estimates. In each round, the software also required participants to provide their individual numerical estimates as well as 10% lower and upper confidence bounds. Furthermore, participants had the possibility to reveal comments. After everyone responded, the results of the first round were summarized and reported as feedback to participants; this included each individual estimate (along with the confidence interval), a group summary of these estimates (mean, median, and standard deviation), as well as each individual's comments per question. Participants then provided their final individual estimates in a second round. For our analysis, we used the median of the individual estimates of the second round to calculate the group results.

Prediction markets

We used the (partly) free software solution of inklingmarkets.com, which was specifically designed to make participation as easy as possible for non-experienced traders. In addition, since our markets had only a small number of participants, it was necessary to rely on a system using a market maker in order to ensure sufficient liquidity of the market. With Inkling's market maker, participants can buy and sell contracts, and thus reveal information, at any time as they directly interact with the market maker and do not require a human trading partner. That way, the market can elicit information even from a single participant. For further discussion of the software, see Christiansen (2007), who conducted field experiments using Inkling.

Each question was mapped as one contract, which could be traded by participants. At market closure, each contract was liquidated at the value of the correct answer for each question. Thus, if one thought the correct answer was higher than the contract's current value, one would have bought shares, otherwise one would have sold. The procedure is explained in detail in a video tutorial, which was watched by each participant before logging into the system (http://tinyurl.com/method-comparison-PM). Each participant had an initial deposit value of $30,000 of play-money. At the end of the session, the last traded contract prices were interpreted as the final group results.

Quantitative judgment task

The questions were influenced by two limitations of the prediction market software. First, although prediction markets can theoretically be used to predict large numbers, it would have been necessary to scale the index markets, which might have been harder to understand for participants. Second, for market maker mechanisms it is necessary to provide a starting price that should reasonably be chosen greater than zero. However, announcing starting prices provides anchors to traders.

To circumvent these problems, the quantitative judgment task consisted of ten almanac questions that required percentage estimates. That way, no scaling was necessary since market prices would always be between 0 and 100. Furthermore, all questions were designed to be similar by providing an anchor in the formulation of the question, which was used as the starting price in the markets. Examples include: "In 1900, the percentage of the total US population that was aged 65 and over was 4.1 %. What was the percentage in 2000?" or "In 1980, 17.0% of the US population (25 years and over) had completed 4 years of college or more. What was the percentage in 2006?" Appendix 2 provides the complete list of questions used in this study.

It was emphasized in the written instructions and by the incentive mechanism that the purpose of the task was to estimate the answers to the questions. Answers were not revealed until completion of the study.

Results

We used the absolute error as an index of accuracy. Tests of statistical significance are reported in footnotes.[2] Significance levels are indicated in tables using asterisks (*: $p \leq 0.05$; **: $p \leq 0.01$). The data that has been analyzed in this study can be found online at http://tinyurl.com/method-comparison-data.

For NGT, Delphi, and six prediction market groups, we had access to participants' ex ante estimates prior to entering group interaction. We used these individual priors as well as the mean of these estimates from the participants in each group as an additional benchmark for our analysis. In the following, the mean prior estimates per group are referred to as *staticized groups*.

Group accuracy before group interaction

We analyzed the accuracy of the staticized groups before group interaction took place. The goal of this exercise was to rule out that differences between the methods might have occurred because some groups simply happened to be better than others. For each question, as well as over all ten questions, we did not find any differences in accuracy between participants' priors in NGT, Delphi, and prediction markets.[3]

Accuracy of structured group interaction vs. individual priors

Table 1 shows the results, reported as the error reduction for each question as well as over all ten questions. The error reduction is the difference between the absolute errors of the prior individual estimates and the respective method result. A positive (negative) value for the error reduction means that the method results were more (less) accurate than the individual priors. As expected, overall, the results of each method were more accurate than the individual priors.[4] At the individual question level, prediction markets were more accurate than individual priors for four questions and equally accurate for six questions. NGT was more accurate for eight questions and equally accurate for two questions. Delphi was most accurate, yielding a lower MAE than the respective individual priors for each of the ten questions.

——Table 1 about here——

Accuracy of group interaction vs. staticized groups

Group interaction should yield to more accurate results than those that can be obtained by simply averaging the prior individual estimates within each group. We compared the methods' results to the corresponding staticized groups. The results, again reported as the error reduction, are shown in Table 2.[5]

---

[2] Since, in most cases, the data were not normally distributed, we conducted non-parametric significance tests.
[3] The Kruskal-Wallis test was used to test for differences between the absolute errors of the prior individual estimates of the three structured approaches.
[4] The Wilcoxon rank-sum test was used to test for differences between the absolute errors of the individual priors and the method results. Table 1 reports one-tailed p-values.
[5] The Wilcoxon rank-sum test was used to test for differences between the absolute errors of the staticized groups and the method results. Table 2 reports one-tailed p-values.

On average, the NGT results were significantly more accurate than the staticized groups for three out of ten questions (i.e., 'population of Australia', 'motor vehicle production', and 'population aged 5-'). For the question 'households with PC', the NGT results were significantly less accurate. For the remaining six questions, there were no statistically significant differences in accuracy.

The Delphi method yielded significantly more accurate results than the staticized groups for three out of the ten questions. For the remaining seven questions, differences were small.

Prediction markets improved on the staticized groups for two questions. For the remaining questions, there were no statistically significant differences, which might be due to the small number of observations. However, although differences were small, prediction markets were less accurate than the staticized groups for six of the ten questions.

Over all ten questions, group interaction improved on the staticized groups. That said, the gains in accuracy were modest and the sample sizes were not large, so there is some uncertainty about these findings. The differences were statistically significant only with respect to Delphi.

——Table 2 about here——

Relative accuracy of group methods

For each method, we calculated the MAEs for each question as well as over all ten questions. With the lowest overall MAE of 5.62, Delphi was most accurate, followed by NGT (5.92) and prediction markets (7.07). As expected, FTF was least accurate; its overall MAE was 7.21. Although the results conformed to our expectations that the structured approaches would outperform FTF, the overall differences between the methods were not statistically significant.[6] On the question level, the relative performance of methods was mixed. Table 3 shows the error reduction of NGT, Delphi, and prediction markets compared to FTF.

——Table 3 about here—

Differences between FTF and NGT were small.[7] For eight of the ten questions, there were no significant differences between the two methods. For the two remaining questions, NGT was once more – and once less – accurate than FTF.

Delphi was significantly more accurate than FTF for two questions. For the remaining eight questions, there were no differences in accuracy. In no case was Delphi significantly less accurate than those from FTF.

[6] The Kruskal-Wallis-test showed no statistically significant differences between the four methods over all 10 questions.

[7] We conducted Mann-Whitney U-tests to compare the results of FTF and the respective structured approaches for each question.

Contrary to our expectations, prediction markets were unable to improve on the accuracy of FTF. Instead, for three questions, prediction markets yielded results that were significantly less accurate than the respective FTF results. For the remaining questions, there were no statistically significant differences.

Participants' perceptions

After the group interaction, we asked participants about their perceptions of the group as well as the group process. We had hypothesized that the personal interaction between participants would lead to more favorable ratings for FTF and NGT compared to Delphi and prediction markets. Furthermore, we had assumed that, due to less group pressures, NGT would be rated more favorable than FTF. Finally, because of participants' unfamiliarity with the approach, we expected prediction markets to be rated most unfavorable on most attributes. The results were consistent with these hypotheses. For each attribute, Figure 1 reports the statistical mean (and 95% confidence intervals) of participants' ratings, revealed on a 7-point Likert scale from "1" (very low) to "7" (very high). Note that for the attributes *disagreement* and *difficulty,* higher ratings were interpreted as negative.

——Figure 1 about here——

Ratings of the group

NGT obtained most favorable ratings for *cooperation* and *disagreement* and ranked second for participants' *confidence* in the group answers. Prediction markets ranked second worst for *disagreement* and *confidence* and worst for *cooperation*. For both *cooperation* and *disagreement*, FTF ranked second.

Although Delphi obtained the worst score for *disagreement*, participants' were most confident in the outcomes. We see the following reasons for the high perceived disagreement in Delphi: Delphi participants cannot communicate directly with each other but reveal their estimates independently. Agreement can only increase if incorporating feedback information after the first round leads to more cohesive results at the end of the process. However, disclosing disagreement can be desirable as it alerts decision-makers to uncertainty.

FTF obtained the lowest score in terms of *confidence* in the group outcome. This is surprising since we would have expected that high levels of perceived cooperation and agreement would result in high levels of confidence in the group results. We assume that FTF participants were aware that, due to group tendency effects, probably not all available information was aggregated from group members.

Ratings of the group process

Rankings on the group process as a whole were identical over 4 out of 5 attributes with NGT placed first, FTF second, Delphi third and prediction markets fourth. For *difficulty to participate*, FTF achieved an equally favorable score as NGT. For *freedom to participate*, FTF and Delphi swapped places. With

respect to favorable ratings for FTF on most other attributes, it is interesting that Delphi achieved a higher score than FTF for *freedom to participate*. Again, the reason might be that FTF participants experienced group pressures that may have hindered them from fully revealing their information, which conforms to the low confidence scores for FTF. Overall, participants' perceptions of the prediction markets process were poor for all attributes.

In sum, the results conformed to our expectations. NGT obtained most favorable ratings for 7 of the 8 categories. As hypothesized, FTF were rated quite favorable, too; ranking second for six attributes. Finally, the results corroborated our hypothesis that prediction markets would be rated most unfavorable; they had the worst ratings for six attributes and were second worst for the two remaining attributes.

Discussion

Although the three structured approaches yielded a lower MAE than FTF over all ten questions, the differences in accuracy between the four methods were not statistically significant. The results conformed to our hypotheses only with respect to the direction of the results.

However, some differences between the methods were identified on the question level. Delphi was the only approach that was never less accurate than FTF and outperformed FTF for two of the ten questions. By comparison, prediction markets were less accurate than FTF for three of the ten questions and were unable to outperform FTF for any of the ten questions. The relative performance of NGT and FTF was mixed and differences were small. We can only speculate on the reasons for these results.

Participants in this study had to solve a quantitative judgment task that required percentage estimates on factual questions. Such tasks have clear solutions, lack highly distributed knowledge and uncertainty, and require groups only to aggregate factual knowledge. We would expect that structured approaches would be substantially more effective when participants reveal information that is new and useful to others.

For example, for site location problems, different experts might have distinct knowledge about labor and real estate costs, traffic patterns, and the types of consumers that live near a potential retail outlet. Other tasks like forecasting, decision-making, negotiation, or conflict solving are more challenging as they not only involve 'facts' but also values, emotions, and expectations, as well as high levels of uncertainty. We would expect the relative performance of methods to be largely influenced by such problems (Wright & Ayton 1986). In particular, we would expect group pressures to be much higher and, therefore, FTF to perform generally worse.

In contrast, when all the experts draw upon the same information, which might have been the case for the type of problem used in this study, information exchange is expected to be of little value. Over all ten questions, only Delphi yielded more accurate results (in terms of statistical significance) compared to

staticized groups. By comparison, differences of prediction markets and NGT compared to staticized groups were not statistically significant.

Under most circumstances, structured combined forecasts are substantially more accurate than individual forecasts (Armstrong 2001). Our results provided further support for this well-established principle. Each of the three structured approaches yielded lower mean absolute errors than the prior individual estimates.

Another factor that might have influenced the results might be the experiment design. The experiment setting favored meetings since it did not involve hierarchies. In a more realistic environment with hierarchies, we would expect group pressures to be a concern. As a result, the relative performance of FTF would decrease.

The comparably poor performance of prediction markets may have been due to limited means of communication between participants. Although participants could exchange information continuously through trading, they were unable to provide reasoning for their estimates. Participants could only buy and sell contracts, without indicating why they did so. Thus, the trading mechanism might not have allowed participants to convey information to the group. However, prediction markets could be designed to overcome such barriers. For example, adding additional means of communication like forums or chat rooms could be implemented that allow traders to communicate with each other and to reveal reasoning for their actions. Although several commercial prediction markets have implemented such solutions, we are unaware of any research that examines whether this has an effect on accuracy.

Examining participants' perceptions of the group and the group process as a whole revealed a preference for methods involving personal communication. These results supported our hypothesis that personal interaction in non-hierarchical meetings can lead to coherence and, therefore, increase perceived satisfaction of participants. In contrast, Delphi and prediction markets were rated less favorable. In particular, prediction markets were rated worst on 6 attributes and second worst on the remaining two. We see at least two reasons: First, the lack of exchanging reasons for judgment may alienate traders. We would expect convenience with the method to rise with increasing insight into the opinions of fellow traders. Second, prediction markets are still a relatively new approach and the idea of revealing one's opinion through trading stocks may be difficult to understand for participants. Even though we relied on software that was specifically designed to make participation as easy as possible for novices, participation in prediction markets was rated by far most difficult. We assume that high perceived difficulty was crucial for the poor ratings on other categories which, in turn, may hinder a further adoption of prediction markets by practitioners. Researchers and developers should further increase their efforts to make market software solutions more accessible to non-experienced traders.

Participants' perceptions of the group process can be crucial for the acceptance of its results. When agreement among – and satisfaction of – group members is high, decision-makers can share

responsibility for their decisions. Thus, our results may entice decision-makers to rely on methods involving personal communication. However, such a strategy can bring along drawbacks. (1) There is no evidence that high perceived satisfaction is correlated to good performance. (2) Taking into account administrative and time efforts, meetings are expensive and can be difficult to conduct, as they require participants' presence. (3) Meetings are limited in the number of participants. In contrast, Delphi and prediction markets do not require presence of participants and can be conducted asymmetrically with a – virtually unlimited – number of people. (4) Meetings might lead to poor decisions. Due to higher group pressures in real world situations, we would expect meetings to perform worse for more complex problems and more realistic environments, especially when relevant information is distributed among the experts.

Conclusion

We compared the relative accuracy of traditional FTF to three structured approaches (NGT, Delphi, and prediction markets) on a quantitative judgment task that required percentage estimates for ten factual questions.

Over all ten questions, we did not find statistically significant differences in accuracy between the four methods. However, the method results somewhat differed at the individual question level. Delphi was as accurate as FTF for eight questions and outperformed FTF for two questions. By comparison, prediction markets were unable to outperform FTF for any of the ten questions but were inferior for three questions. The relative performance of NGT and FTF was mixed and differences were generally small.

The three structured approaches were more accurate than participants' prior individual estimates. Delphi was also more accurate than staticized groups, in contrast to NGT and prediction markets. The results suggest that NGT and prediction markets provide little additional value compared to simple averages of forecasts in situations where participants draw upon similar information. Future research should evaluate the relative performance of the methods for more complex problems in more realistic environments.

We also analyzed participants' perceptions of the group and the group process as a whole. Participants rated methods involving personal communication (FTF and NGT) more favorable than the computer-mediated Delphi and prediction markets. In particular, FTF and NGT participants experienced higher levels of cooperation among their groups and perceived group interaction as more effective. Prediction markets were rated least favorable. Prediction market participants were least satisfied with the group process and rated their method highest in terms of difficulty of participation.

Acknowledgments

We thank Kesten Green, Robin Hanson, and David Pennock for helpful comments. Thanks to the staff of the Wharton Behavioral Lab, in particular Daniela Lejtneker, Tatiana Silva, and Patricia Zapater-roig. Ellen Rosenberg and Christian Menn helped with preparing materials. Thanks to inklingmarkets.com for providing the prediction markets software.

References

Armstrong, J. S. (2001). Combining forecasts. In: J. S. Armstrong (Eds.), *Principles of Forecasting. A Handbook for Researchers and Practitioners.* Boston, MA: Kluwer Academic Publishers, pp. 417-439.

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings, *Foresight - The International Journal of Applied Forecasting*, Issue 5, 3-8.

Berg, J. E., Nelson, F. D. & Rietz, T. A. (2008). Prediction market accuracy in the long run, *International Journal of Forecasting*, 24, 285-300.

Boje, D. M. & Murnighan, J. K. (1982). Group confidence pressures in iterative decisions, *Management Science*, 28, 1187-1196.

Christiansen, J. D. (2007). Prediction markets: Practical experiments in small markets and behaviours observed, *Journal of Prediction Markets*, 1, 17-41.

Erikson, R. S. & Wlezien, C. (2008). Are political markets really superior to polls as election predictors?, *Public Opinion Quarterly*, 72, 190-215.

Fischer, G. W. (1981). When oracles fail--a comparison of four procedures for aggregating subjective probability forecasts, *Organizational Behavior and Human Performance*, 28, 96-110.

Green, K. C., Armstrong, J. S. & Graefe, A. (2007). Methods to elicit forecasts from groups: Delphi and prediction markets compared, *Foresight - The International Journal of Applied Forecasting*, Issue 8, 17-20.

Gustafson, D. H., Shukla, R. K., Delbecq, A. & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups, *Organizational Behavior and Human Performance*, 9, 280–291.

Rhode, P. W. & Strumpf, K. S. (2004). Historical presidential betting markets, *Journal of Economic Perspectives*, 18, 127-141.

Rosenbloom, E. S. & Notz, W. (2006). Statistical tests of real-money versus play-money prediction markets, *Electronic Markets*, 16, 63-69.

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

Servan-Schreiber, E., Wolfers, J., Pennock, D. M. & Galebach, B. (2004). Prediction markets: Does money matter? *Electronic Markets*, 14, 243 - 251.

Van de Ven, A. H. & Delbecq, A. L. (1971). Nominal versus interacting group processes for committee decision making effectiveness, *Academy of Management Journal*, 14, 203-212.

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

Wolfers, J. & Zitzewitz, E. (2004). Prediction markets, *Journal of Economic Perspectives*, 18, 107-126.

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

Wright, G. & Ayton, P. (1986). Subjective confidence in forecasts: A response to Fischhoff and MacGregor, *Journal of Forecasting*, 5, 117.

Table 1: Mean error reduction of each method compared to prior individual estimates before group interaction

	NGT	Delphi	PM
Question	(N=54)	(N=59)	(N=34)
Population aged 65+	1.33*	3.15**	0.88
Internet users	2.69*	3.50**	1.84
Health expenditures	3.60**	2.99**	0.85
College graduates	4.12*	4.95**	6.51**
Population of Australia	2.42**	1.48**	2.67**
Overweight children	1.77	0.79**	0.79
Children with two parents	2.57**	4.51**	4.06**
Households with PC	0.78	5.08**	3.09
Motor vehicle production	9.82**	5.29**	5.43
Population aged 5-	1.86**	1.24**	4.42**
Overall	**3.10****	**3.60****	**3.05****

N: number of prior individual estimates

Table 2: Mean error reduction of each method compared to staticized groups

	NGT	Delphi	PM
Question	(N=11)	(N=11)	(N=6)
Population aged 65+	-0.87	0.93	-1.7
Internet users	0.89	1.59	-1.13
Health expenditures	0.71	0.09	-3.22
College graduates	0.3	1.52	3.25
Population of Australia	2.13**	1.45**	2.47*
Overweight children	-2.77	-1.02	-3.52
Children with two parents	-1.69	-1.39	-1.37
Households with PC	-4.96**	-2.09	-1.37
Motor vehicle production	8.79**	3.58*	3.95
Population aged 5-	1.58**	0.73*	3.77*
Overall	**0.41**	**0.54***	**0.17**

N: number of groups

Table 3: Error reduction of NGT, Delphi, and prediction markets compared to FTF

	MAE	Error reduction to FTF		
Question	**FTF**	**NGT**	**Delphi**	**PM**
Population aged 65+	7.76	2.69	4.29**	1.22
Internet users	4.53	-0.53	-1.33	-3.36
Health expenditures	3.86	2.19	1.38	0.16
College graduates	6.36	0.27	-0.26	1.87
Population of Australia	2.41	-0.09	-1.2	-2.35*
Overweight children	13.98	7.93*	9.74**	5.52
Children with two parents	7.19	0.8	2.77	0.98
Households with PC	15	0.82	7.21	6.5
Motor vehicle production	10.12	-0.22	-5.49	-7.92*
Population aged 5-	0.86	-1.00*	-1.21	-1.26**
Overall	**7.21**	**1.29**	**1.59**	**0.14**

Figure 1: Mean participants' perceptions along categories, revealed on a 7-point Likert scale