Mitigating manipulation in committees: Just let them talk!

David Albrecht*

October 10, 2023

[Link to most recent version]

Abstract

Many decisions rest on the collective judgment of small groups like committees, boards, or teams. However, some group members may have hidden agendas and manipulate this judgment to sway decisions in their favor. Utilizing an incentivized experiment, I analyze how manipulation affects the objective accuracy and perceived trustworthiness of such group judgments depending on the format of group interaction.

I compare group judgments from unstructured face-to-face interactions, which are ubiquitous in real-world institutions, to group judgments from the scientifically endorsed, structured Delphi technique. To identify mechanisms underlying the accuracy differences, I use structural estimations and analyze emergent communication patterns.

Without manipulation, Delphi is more accurate than face-to-face interaction and indistinguishable from the Bayesian benchmark. Manipulation decreases accuracy for Delphi but not for face-to-face interaction. Thus, with manipulation, Delphi is less accurate than face-to-face interaction. With manipulation, sharing of (truthful) information decreases in Delphi but not in face-to-face interaction. The structural estimations further reveal that Delphi judgments likely exhibit more bias towards hidden agendas and less utilization of valuable information.

Perceived trustworthiness does not always match objective accuracy. Judgments from face-to-face interaction - unjustifiably - enjoy higher levels of trust without hidden agendas. Trustworthiness correctly decreases with hidden agendas for Delphi groups but - unjustifiably - also for face-to-face groups. With hidden agendas, face-to-face groups are simultaneously more accurate and trusted.

Keywords Committees · Manipulation · Face-to-face Communication · Delphi Technique

JEL Classification D70, D71, D83

1 Introduction

In many institutions, high-stake decisions are made based on the collective judgment of small groups, like (expert) committees or advisory boards that evaluate some decision-relevant criteria. Examples of such group judgments span diverse contexts: scientists evaluate the effectiveness of policy interventions to fight global pandemics (Haug et al., 2020), managers rate investment alternatives (Lovallo et al., 2020), and security experts assess terrorist activity (Friedman and Zeckhauser, 2014).

The question remains whether the widespread use of this practice implies its appropriateness. To optimally inform decisions, the accuracy and trustworthiness of group judgments are essential. However, some group members may have hidden agendas and manipulate this judgment in a particular direction to sway consequent decisions in their favor.[1] Generating accurate and trusted collective intelligence in such a setting remains a challenge. Therefore, I analyze whether two common formats of group interaction may preserve accuracy and trustworthiness despite manipulation due to hidden agendas. In particular, I compare group judgments from unstructured face-to-face interactions, which are ubiquitous in real-world institutions, to group judgments from the scientifically endorsed, structured Delphi technique.

A way of extracting collective intelligence from groups that is both accurate and perceived as trustworthy has yet to be identified. Previous research in this context has focused on prediction markets and exogenously structured face-to-face interactions. While prediction markets appear accurate but not always trusted, structured face-to-face interactions are highly trusted but not always accurate. Prediction markets proved themselves as a robust tool to accurately elicit and aggregate dispersed information from a group of people (Wolfers and Zitzewitz, 2004; Arrow et al., 2008), which has also been successfully tested by large institutions such as Google and Ford (Cowgill and Zitzewitz, 2015). Such markets remain informative even if some participants act manipulatively. However, in

---

[1]E.g., serving their career in the institution (Toma and Butera, 2009; Mattozzi and Nakaguma, 2022), serving other organizational units the individual is associated with (Pearsall and Venkataramani, 2015), benefiting their reputation (Visser and Swank, 2007), overstating chances of the political victory of favored parties (Hansen et al., 2004; Rothschild and Sethi, 2016), receiving bribes (Felgenhauer and Grüner, 2008) or downplaying security threat to reduce responsibility and effort for consequent duties (Amjahid et al., 2017).

response to manipulation, decision-makers observing the market lose trust and make even worse decisions, as if they had ignored the market completely (Kaplan et al., 2022).

Face-to-face interactions are a standard operating procedure in institutional decision-making. The particular variety of such interactions, that has been studied in hidden agenda settings, are nominal groups. These are groups interacting in an exogenously imposed estimate-talk-estimate format and have been found to enjoy high levels of trust among decision-makers.[2] Adversely, trustworthiness persists even if judgment accuracy deteriorates with manipulation (Maciejovsky and Budescu, 2020). Taken together, both prediction markets and nominal groups appear suboptimal and may lead to many objectively ill-informed decisions with high societal stakes. To address this problem, I extend previous investigations to two alternative formats of group interaction. On the one hand, I analyze commonly used face-to-face interaction (FTF) that is not exogenously bound to any structure, such as the protocol of nominal groups. On the other hand, I consider the scientifically endorsed Delphi technique (Delphi), where interaction is exogenously structured according to a specific protocol.[3] I quantify the extent to which these interaction formats produce a collective intelligence that is simultaneously accurate and perceived as trustworthy. Further, I identify what drives robustness towards manipulation.

To measure the capabilities of Delphi and FTF, I developed a pre-registered, incentivized lab experiment.[4,5] that enables the isolation of the causal effects of hidden agendas and the format of group interaction on quantitative measures of objective accuracy and perceived trustworthiness. Conducting a lab experiment offers a unique opportunity to address the research questions by controlling for many factors that are not directly observable in real-world situations where group judgments are used. Hidden agendas, by their

---

[2]In nominal groups, group members first estimate individually, second talk with their group members, and finally estimate individually again.

[3]FTF and more broadly free-form deliberation of groups are ubiquitous in real-world institutional decision-making (Maciejovsky and Budescu, 2020) despite concerns about their performance (Sunstein, 2005; Armstrong, 2006). Though less common in practice, the Delphi technique is a group interaction format supported by empirical and theoretical research (Rowe and Wright, 1999; Graefe and Armstrong, 2011).

[4]Pre-registration files can be found via OSF https://osf.io/cv4md/

[5]The experiment was reviewed and approved by the Ethical Review Committee Inner City Faculties at Maastricht University (reference ERCIC_267_10_06_2021).

very nature, are covert to researchers. Moreover, group judgments potentially depend on many, not directly measurable factors, such as the expertise of group members, and are used in situations where the true answer is unknown.

The experiment comprises two parts: a group *estimation experiment* and a subsequent *decision experiment.* In the first part, the *estimation experiment*, I examine the objective accuracy of groups of four people who collaborate on a series of probabilistic judgment tasks in which they generate joint group judgments. The experiment follows a two-by-two between-subject design with FTF and Delphi groups, each with and without hidden agendas. In the FTF treatments, groups interact via face-to-face conversation in a video call without any imposed structure. In the Delphi treatments, groups interact through an imposed, pseudonymous chat protocol. According to this protocol, group members first make individual quantitative and qualitative judgments, then review all members' pseudonymous judgments and reasonings before producing a second quantitative estimate individually. Across all treatments, I incentivize accuracy at the group level. In the hidden agenda treatments, I additionally induce manipulation incentives (hidden agendas) in half of the group members through individual, private side payments.

In the second part, the *decision experiment*, I elicit the perceived trustworthiness of the group judgments obtained from the estimation experiment. A new set of individual participants who did not participate in the estimation experiment states their individual, incentivized confidence intervals around group judgments.[6] Each participant evaluates judgments from all four treatment conditions. This decision experiment allows for comparisons of the perceived trustworthiness of group judgments produced by FTF versus Delphi groups, each with and without hidden agendas, respectively.

I identify differences in objective accuracy by comparing absolute errors across conditions of the estimation experiment. Without hidden agendas, structured Delphi interaction produces more accurate judgments than FTF. This result reverses with manipulation, which does hamper accuracy in Delphi groups but not in FTF groups. To put accuracy into perspective, I compare Delphi and FTF groups against the best Bayesian benchmark based on the information available to the groups. Without hidden agendas, the best benchmark's accuracy is statistically indistinguishable from that of Delphi but sig-

---

[6]Intervals are incentivized through the incentive-compatible most likely interval technique (Schlag and van der Weele, 2015).

nificantly more accurate than FTF. Both interaction formats, Delphi and FTF, perform significantly better than a benchmark, which estimates all probabilities naïvely at 50%. In situations with hidden agendas, no interaction format reaches the best Bayesian benchmark. While FTF groups still outperform the naïve heuristic, Delphi falls behind this benchmark.

I identify differences in the perceived trustworthiness of group judgments by comparing the length of stated confidence intervals in the decision experiment across conditions of the estimation experiment. FTF group judgments are generally trusted more than Delphi group judgments. This holds regardless of the presence of hidden agendas and also in cases where it is not justified by objective accuracy.

All in all, FTF group judgments appear problematic in situations without hidden agendas. Here, they are objectively less accurate. Decision-makers fail to realize this and put relatively more trust in FTF judgments. This can be taken as evidence that FTF should be used less in institutional decision-making in situations without hidden agendas, which are arguably very rare. The Delphi technique seems to be a better alternative. However, in situations with hidden agendas, perceived trustworthiness is aligned with objective accuracy, as the relatively more accurate judgments from FTF groups are also trusted more than judgments from Delphi groups. This can be taken as evidence that FTF is an appropriate format for institutional decision-making in situations with hidden agendas, which are arguably very prevalent.

To explore the mechanisms underlying the differences in accuracy, I use structural estimations and analyze emergent communication patterns across conditions of the estimation experiment. In particular, I analyze the data from group estimation in the BIN-modeling framework (Satopää et al., 2021) (Section 4.4.1). This allows me to decompose the estimation error in all conditions of the estimation experiment into: (i) bias (consistently producing too high or too low estimates), (ii) noise (coming to different conclusions given the same information), and (iii) the usage of valuable information. Delphi, as compared to FTF, generally exhibits less noise. However, with the introduction of hidden agendas, Delphi likely suffers from more bias, while FTF appears robust against this effect.

Complementing the results from the structural estimations, I analyze the degree and truthfulness of information sharing during group interaction. Transcribed and coded communication protocols of FTF and Delphi interactions reveal that hidden agendas lead

to less information sharing and decrease the truthfulness of shared information for Delphi groups. By contrast, in FTF groups, the amount of shared, truthful information increases with hidden agendas.

Condensed, FTF seems generally better suited for group judgments in situations with hidden agendas. Moreover, FTF appears to be the better choice in situations without hidden agendas where the perceived trustworthiness of group judgments is highly prioritized. The Delphi technique is preferable in situations without hidden agendas where accuracy is the highest priority.

2   Related literature

This work connects to the broad interdisciplinary literature on collective intelligence. The most important links are to comparisons of multiple group interaction formats, investigations on single group interaction formats in isolation, and the effects of manipulation or lying behavior on group judgments. Furthermore, the studied setting is related to the longstanding research on committee decision-making.

This study is closest to **comparisons of group interaction formats** concerning accuracy, trustworthiness, and robustness towards manipulation. Early studies in this area focus on comparing the accuracy of the Delphi technique to other formats of structured and unstructured direct interaction, as well as averaging individual judgments. This literature attests to mixed results on the relative performance of the Delphi technique (Woudenberg, 1991; Rowe and Wright, 1999). However, reviews also acknowledge severe methodological limitations in early Delphi research. Notably, many studies used (over)simplified Delphi designs, which might have led to Delphi's capacity being understated (Rowe et al., 1991). Later, prediction markets likewise moved into the focus of collective intelligence research, and advances in software allowed more consistent implementation of structured interaction formats in the laboratory. Computerized Delphi groups rank most accurate in a comparison against unstructured face-to-face groups, structured nominal groups, and groups trading in prediction markets to solve general knowledge questions in the laboratory (Graefe and Armstrong, 2011). Based on data from a geopolitical forecasting tournament, prediction markets outperform median estimates of interacting teams but lose to more sophisticated statistical aggregation (Atanasov et al., 2017). Previously mentioned studies consider situations without the explicit threat

of manipulation. Maciejovsky and Budescu (2013) are the first to consider conflicts of interest that induce manipulation in structured nominal groups vs. prediction markets in a lab experiment. In contrast to nominal groups, prediction markets maintain knowledge sharing if the existence of conflicts of interest is commonly known. In a follow-up, Maciejovsky and Budescu (2020) additionally consider the trustworthiness of judgments. Without manipulation, nominal groups are more accurate than markets, but markets are more accurate than groups with manipulation. At odds with accuracy, nominal groups' judgment is perceived as more trustworthy overall, especially in the manipulation treatments. This holds for trustworthiness as evaluated by group members themselves but also by external observers. Similarly, Kaplan et al. (2022) compare the accuracy of prediction markets without manipulative traders to markets where manipulators are active with a certain probability and markets where manipulators are active with certainty. The mere potential of manipulation hampers accuracy. Nevertheless, even with certain manipulation, markets still reveal valuable information. For decision-makers observing the market, already the potential of manipulation and, more so, certain manipulation causes erosion of trust. Remarkably, while markets with manipulation still reveal objectively valuable information, erosion of trust prevents decision-makers from utilizing it. In fact, with certain manipulation, decisions are even worse, as if decision-makers would have ignored the market entirely.

Extracting collective intelligence from groups in hidden agenda settings remains a challenge. Previous research, as summarized in Table 1, has not yet identified interaction formats that are simultaneously accurate and trusted. I built on and extend this body of literature by scrutinizing the accuracy and trustworthiness of group judgments from Delphi groups as opposed to freely interacting face-to-face groups. To the best of my knowledge, I am first to compare these two interaction formats in a setting with hidden agendas. Further, I advance the methodology developed in previous method comparisons by combining experimentally induced, complementary expertise, and continuous as well as incentivized measures of accuracy and trustworthiness.

Focusing on a single group interaction format at a time, **manipulation** in collective intelligence has further been studied for prediction markets. Evidence is mixed with some studies reporting manipulation to harm accuracy (Hansen et al., 2004; Gimpel and Teschner, 2014; Kaplan et al., 2022), while others find prediction markets to be robust

Table 1: Comparisons of group judgment techniques in the literature

	Graefe and Armstrong (2011)	Maciejovsky and Budescu (2013)	Maciejovsky and Budescu (2020)	Kaplan et al. (2022)	This study
Interaction format (# of group members in parentheses)					
unstructured face-to-face	✓(4-6)	-	-	-	✓(4)
Delphi technique	✓(4-6)	-	-	-	✓(4)
nominal group technique	✓(3-6)	✓(3)	✓(3)	-	-
prediction markets	✓(4-7)	✓(3)	✓(3)	✓(8)	-
Assessed quality measures					
accuracy	✓	✓	✓	✓	✓
trustworthiness	-	-	✓	✓	✓
Features of the group judgment task					
hidden agendas	-	✓	✓	✓	✓
induced expertise	-	✓	✓	✓	✓
Sample					
n judgment	227	144	450	112	590
groups/ condition	11	16	15	7	15
n trust	-	-	224 + 358	56	2000

towards manipulation (Hanson et al., 2006; Hanson and Oprea, 2009; Teschner et al., 2017). Beyond markets, the Delphi technique appears manipulable by the Delphi administrators (Nelson, 1978), and Delphi estimates may generally be biased by the desirability of outcomes to Delphi group members Ecken et al. (2011). Further, (Wittrock, 2023) discusses theoretically that Delphi group members may not report truthfully in order to affect participants and ultimately influence decisions tied to the Delphi result. Yet, explicit manipulation of Delphi estimates by group members has not been empirically

investigated prior to the present study.

To study the influence of manipulation, this work strongly builds on prior research on the **Delphi technique**, a structured interaction format to extract and aggregate human judgment from a group of experts. The format aims to strengthen positive aspects of deliberation while minimizing process loss of group interaction (Rowe and Wright, 2001). Two basic concepts build the foundation of Delphi's potential. First, according to the theory of errors, combining estimates of multiple judges will lead to a group judgment that is more accurate than the average individual's judgment (Dalkey, 1975). Second, throughout the interaction, relatively less accurate judges will update more than relatively accurate judges (Parenté and Anderson-Parenté, 1987), and relatively more accurate feedback has a stronger influence on such updating (Rowe et al., 2005; Bolger and Wright, 2011). Both concepts are also noted in the broader context of the "wisdom of crowds", i.e., aggregating independent, individual judgments is more accurate than the average individual (Galton, 1907). Second, through information exchange, individual judgments lose independence. However, accuracy may still increase (Mellers et al., 2014; Da and Huang, 2020), and combined group judgments improve if relatively stronger influence is exerted from more to less accurate individuals (Becker et al., 2017).

The Delphi technique dates back to research by the RAND Corporation in the late 1940s to improve expert forecasting on topics such as technological change. Subsequently, manifold research projects used the Delphi technique and investigated its performance and workings as, e.g., reviewed by Rowe et al. (1991); Woudenberg (1991); Rowe and Wright (1996, 1999); Bolger and Wright (2011). In parallel, research from various disciplines has used the Delphi technique to generate estimates on transcontextual topics such as climate change mitigation measures (Griscom et al., 2017), biotechnologies improving health in developing countries (Daar et al., 2002) and geopolitical events (Wintle et al., 2012). Implementations of the Delphi technique often differ in their specific design. However, four key characteristics manifest: anonymity, controlled feedback, iteration, and statistical aggregation of final inputs (Rowe and Wright, 2001; Grime and Wright, 2016; Belton et al., 2019). Moreover, certain Delphi features empirically appear accuracy enhancing: feedback that also includes written rationales next to group members' estimates (Best, 1974; Rowe and Wright, 1996; Rowe et al., 2005; Bolger and Wright, 2011; Bolger et al., 2011) and group members with objective (complementary) expertise (Jolson and Rossow, 1971; Rowe

and Wright, 1996). The present study incorporates Delphi's four key characteristics, feedback of written rationales, and experimentally induced, complementary expertise into implementing Delphi interaction in the estimation experiment.

In a broader context, this study builds on the literature on **committee and jury decision-making** going back to (Condorcet, 1785). A common focus of this research strand lies on a group of people, i.e., the jury or committee, who reach a joint verdict, e.g., convict or acquit, by some voting procedure, which may be preceded by deliberation. Some jury members may follow strategic motives as, e.g., discussed by Feddersen and Pesendorfer (1998). This has common features with the given setting of groups deriving a joint probabilistic judgment but also exhibits apparent differences, e.g., the given setting does not explicitly study specific voting rules, nor is a decision made by the committee itself. Notably, a phase of communication preceding voting was found to counteract hidden agendas by inducing mostly truthful information sharing and less manipulative voting (Goeree and Yariv, 2011). However, making such deliberation public to non-committee members counteracts these positive effects (Fehrler and Hughes, 2018). Moreover, if committee members differ in competence, the voting behavior's transparency may counteract competent members' hidden agendas (Mattozzi and Nakaguma, 2022).

Furthermore, the effects of manipulation on collective intelligence strongly relate to the general literature on **lying and deception**. In contrast to predictions of standard economic theory, many individuals tell the truth despite potential monetary gains from lying (Gneezy, 2005; Sutter, 2009). Further, many of those who do not tell the truth do not lie to the full extent (Mazar et al., 2008; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018), obfuscate the truth in vague messages (Serra-Garcia et al., 2011) or evasive lies (Khalmetski et al., 2017). The most plausible explanations for such (non-)lying behavior are preferences for being honest and being seen as honest (Abeler et al., 2019). Lying however increases if individual responsibility is diffused, such as in team settings (Conrads et al., 2013), if the social distance between a liar and the person being lied to increases (Hermann and Ostermaier, 2018), if communication is computer-mediated (Marett and George, 2012) and if the truth that is twisted in a lie, is hard to observe by others (Fries et al., 2021; Hermann and Brenig, 2022). Switching perspectives, many people do not correctly detect if lied to (Bond and DePaulo, 2006), are overconfident about their ability to do so, and share undetected lies as truth (Serra-Garcia and Gneezy,

2021). I compare these results to my analysis of (truthful) information revelation during group interaction across conditions of the estimation experiment.

3   Experiment

The study comprises two complementary experiments: the *estimation experiment*, to examine the accuracy of group judgments, and the subsequent *decision experiment*, to elicit the trustworthiness of group judgments. All experiments are programmed in oTree (Chen et al., 2016) and were implemented onsite in the experimental economics laboratory BEElab at Maastricht University in 2022 and 2023. 240 people participated in the estimation experiment, split into 60 groups of four, 15 groups per treatment condition. Subsequently, 50 participants individually evaluated group judgments in the decision experiment. All participants were recruited from the BEElab participant pool via ORSEE (Greiner, 2015) and followed a study program at the time of the experiment or in the past. Their backgrounds span arts, just as well as social and natural sciences. The educational background's clear mode is business and/or economics (79.6% in the estimation experiment and 74% of participants in the decision experiment). Most participants identified as female (56.3% in the estimation experiment and 64% of participants in the decision experiment). Many had at most one year of professional working experience (70.4% in the estimation experiment and 76% of participants in the decision experiment).

3.1   Estimation experiment

To isolate the causal effect of group interaction format and hidden agendas on the accuracy of group judgments, I implement the estimation experiment as a two-by-two between-subject design. In all treatment conditions, groups are asked to jointly generate a series of ten probabilistic group judgments in a group of four. Groups receive incentives for doing so as accurately as possible. They interact either in FTF or Delphi format. In the hidden agenda treatments (HA), half of the group members additionally have private, individual manipulation incentives that conflict with and outweigh the group incentives for accuracy (see Appendix A.1.2 for details on incentives).

The estimation experiment follows the same structure in all four conditions, as illustrated in Figure 1. First, participants receive video instructions on the judgment task and the interaction format, followed by a comprehension check and a task trial to experience

11

Figure 1: Timeline of the estimation experiment
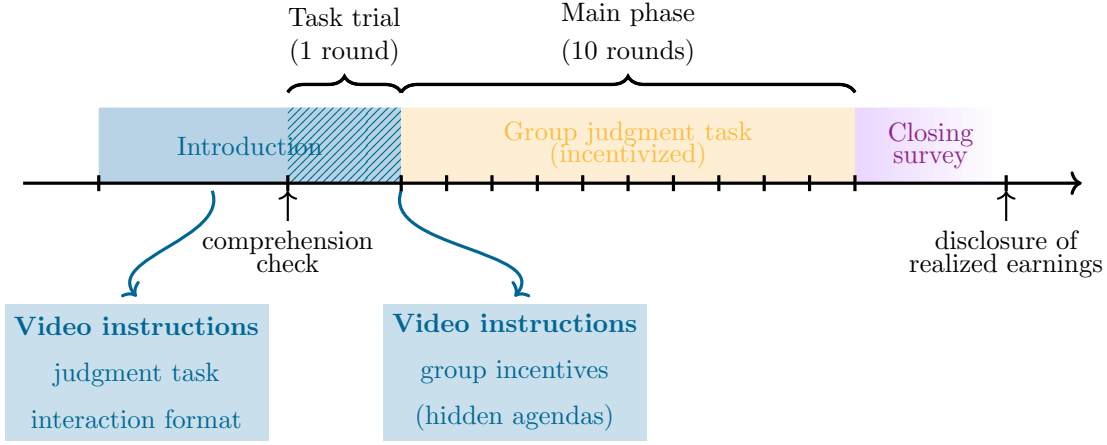
the judgment task and the interaction format. After the trial, before starting the main phase, participants are informed about the incentive scheme and consequently also get to know whether they interact in a setting with hidden agendas. In settings with hidden agendas, participants get to know their type (with or without a hidden agenda) and that exactly two group members have a hidden agenda, and the remaining two have no hidden agenda. Thus, knowledge about one's own type allows one to infer the distribution of types among the remaining group members. Types do not change throughout the whole experiment. Furthermore, in settings with hidden agendas, all participants are informed that hidden agendas may take one of two forms, driving estimates up or down. It is common knowledge that the direction of the hidden agendas is determined randomly per round of the judgment task and that it is the same among the two group members with hidden agendas in any given round. In a specific round, the two group members with a hidden agenda know that their hidden agenda in this round is to drive estimates up (down). This information is not disclosed to the remaining two group members without a hidden agenda. Before starting the main phase, participants are prompted to ask any question to the experimenter that may have remained unanswered.

In the main phase of the experiment, participants go through ten incentivized rounds of solving the probabilistic judgment task. Each round has a distinct underlying, objectively true answer against which the accuracy of group judgments can be evaluated. At the beginning of each round, group members receive individual, complementary information. Combining the individual pieces of information is generally advantageous to generate more accurate group judgments. In groups with hidden agendas, group mem-

bers of hidden agenda type additionally learn the round-specific direction (up or down) of their hidden agenda at the beginning of each round. No information on group accuracy and hidden agenda achievement is presented until the amount of realized earnings is disclosed to each individual at the end of the experiment. Before earnings disclosure and payment, participants complete a questionnaire on socio-demographic characteristics and their subjective perception of the task.

3.1.1 Interaction formats

This paper focuses on the comparison of two interaction formats: free-form face-to-face interaction (FTF) without exogenously imposed structure, which is ubiquitous in real-world institutional decision-making (Maciejovsky and Budescu, 2020), and the Delphi technique, a group interaction format supported by empirical and theoretical research (Rowe and Wright, 1999). In a broader context, the general experimental framework contributes a test bed, which is well suited to investigate the accuracy and trustworthiness of any group interaction format directed at generating quantitative group judgments.

Following Graefe and Armstrong (2011), I implement a very simple form of face-to-face interaction in the **FTF** treatments. Groups are almost unrestricted in their interaction. For each round of the task, group members use a video call in which they may freely discuss the judgment task, their available information, and their strategy to form a joint group judgment.[7] At the end of an interaction, each group member is prompted to enter the joint group estimate for this round. The experiment only proceeds after all group members entered the same estimate, thus enforcing consensus.[8] As such, FTF is not anonymous and allows a free and unrestricted flow of information as well as the full richness of communication (speech, body language, etc.). This resembles typical real-world group interactions (Maciejovsky and Budescu, 2013). For further analysis, the video calls of all face-to-face interactions are recorded and transcribed.

[7]Video-conference interactions have become a very popular medium for teams to interact, latest during the COVID-19 pandemic. Their performance to enable information sharing and to produce group results of high quality appears not different from physical face-to-face interactions (Jabotinsky and Sarel, 2020).

[8]To prevent extraordinarily long discussions, the interaction phase is limited to 10 minutes for the first incentivized round and 7:30 minutes for all later rounds. Groups that do not complete a given round in time do not earn any bonuses on this round. Out of 30 FTF groups, generating a total of 300 group judgments, the interaction on 10 group judgments exceeded the time limit.

In general, **Delphi** evolved to be an umbrella term for diverse forms of structured group interactions. This estimation experiment may only implement and analyze one particular variation of group interaction with Delphi features. To choose a meaningful design for this Delphi interaction, I follow two strategies: First, the design is as simple as possible while still comprising the most common Delphi features: anonymity, controlled feedback, iteration, and statistical aggregation of final inputs (Rowe et al., 1991; Woudenberg, 1991; Grime and Wright, 2016). Second, the design is similar to recent implementations of Delphi procedures in the field of collective intelligence (Becker et al., 2021; Belton et al., 2021).

In the experiment, Delphi groups interact following a standardized, computer-mediated protocol. They do not directly discuss but rather interact through a messaging interface. After receiving their information, group members first provide their numerical, individual estimates and corresponding reasoning in written form. Second, after all group members completed their first input, everyone is presented with the inputs of the three remaining group members and their own input again. This feedback is shown as "the input of Person A, B, and C", and "your input". In this way, linking the estimates and reasonings to a specific real person is impossible. A, B, and C, however, are pseudonyms associated with the same people throughout the experiment. Next, group members are prompted to give a second estimate, which might, but does not have to be a revision of their first estimate, based on the feedback. After all group members completed their second estimate, the group judgment is calculated as the mean of all second estimates. Participants are presented with the group judgment and the underlying, second individual estimates before proceeding with the experiment. Again, this feedback is pseudonymous.

3.1.2 Judgment task

Inspired by (Peeters and Wolk, 2018, 2017), participants are asked to estimate the likelihood that a (biased), two-dimensional, discrete random walk ends above a threshold level after ten steps. Specifically, the random walk starts at 0 and comprises ten steps $X_t = X_1, X_2, ..., X_{10}$. Each step X_t may take the values -1, 0 and 1. Independent draws from an identical distribution determine X_t. The task is repeated for ten rounds. The distribution underlying the random walk varies from round to round of the judgment task. In each round, group members need to estimate $\mathbb{P}(\sum_{t=1}^{10} X_t > 0)$, i.e., the probability that

the random walk takes a value larger than the threshold of 0 after ten steps. Throughout the ten rounds, groups face steps X_t from probability distributions corresponding to $\mathbb{P}(\sum_{t=1}^{10} X_t > 0) = \{0.1, 0.2, 0.3, 0.4, 0.45, 0.55, 0.6, 0.7, 0.8, 0.9\}$ in random order. The task is presented to participants as a computer-generated movement path (random walk) of a ladybird, which may or may not reach a target after ten steps. Groups are asked to estimate the chance that the ladybird reaches the target. They are informed that the computer program generates the movement paths by independently drawing the value per step from an underlying identical distribution, but the exact distribution is undisclosed. To learn about the underlying probabilities, each group member privately receives a movement path generated by a distinct past execution of the computer program. In each round, group members first receive their information and then enter a phase of interaction, which results in a group estimate (see Appendix A.1.1 for further details on the judgment task).

3.2 Decision experiment

This second experiment serves the purpose of evaluating the perceived trustworthiness of estimates generated by experts in the estimation experiment. To this end, experimental participants (decision-makers) who were not involved in the estimation experiment take part in the decision experiment. Within-subject, each decision-maker states their incentivized confidence intervals around 40 estimates generated in the estimation experiment. They face ten randomly selected group judgments drawn from each of the four experimental conditions of the estimation experiment (FTF, FTF + HA, Delphi, Delphi + HA).[9,10] Further, decision-makers are introduced to the judgment task. They are informed about the details of the interaction format and incentives in each of the four treatment conditions in the estimation experiment. Subsequently, decision-makers state the lower and upper bound of an interval for each estimate, which they think likely contains the true

---

[9] This is an adjustment to pre-registration that allows all estimates of the estimation experiment to be evaluated in as balanced as possible: each estimate is evaluated at least three times and by three different decision-makers. The pre-registration outlines that each decision-maker faces all 10 group judgments made by one group in the estimation experiment.

[10] The order in which group judgments from the four experimental conditions of the estimation experiment are evaluated is randomized across decision-makers.

value estimated in the estimation experiment. The width of stated confidence intervals can be seen as a measure of the perceived trustworthiness of group judgments generated in the estimation experiment; the narrower the intervals, the more trustworthy the group judgments, and vice versa.[11]

4 Analysis

4.1 Hypotheses

Initially focusing on the accuracy of group judgments, I test three pre-registered hypotheses. **H1:** Without hidden agendas, Delphi groups are more accurate than FTF groups. **H2:** Hidden agendas impair accuracy in Delphi and FTF groups. **H3:** The negative effect of hidden agendas is relatively smaller for Delphi groups. H1 and H3 relate to the theoretical and empirical arguments that Delphi extracts accurate group judgments by strengthening positive aspects of interaction while minimizing process loss of group interaction (Rowe and Wright, 2001). H2 follows from the intuitive interpretation of hidden agendas as obstacles to accuracy.

Subsequently evaluating the trustworthiness of group judgments, I test three further pre-registered hypotheses. **H4:** Delphi groups enjoy the same level of trust as FTF groups in settings with and without hidden agendas, respectively. **H5:** Hidden agendas impair trust in Delphi and FTF groups. **H6:** The negative effect of hidden agendas on trustworthiness is the same for Delphi and FTF groups. Previous literature has focused on the relative accuracy of Delphi and FTF but remains silent on the relative trustworthiness. As such, H4 and H6 can be seen as a result of an impartial prior about the interaction formats' trustworthiness. H5 follows from the intuitive interpretation of hidden agendas as obstacles to trustworthiness.

4.2 Evaluating accuracy of estimates

To measure the causal effect of hidden agendas and interaction formats on accuracy, I compare group judgments across conditions of the estimation experiment. Based on the estimation experiment, Delphi groups appear more accurate than FTF without hidden

---

[11]Judgments of FTF groups that were NA due to exceeding the time in the experiment (this applies to 10 judgments out of 300) were imputed by 0.5.

agendas, supporting H1. Hidden agendas impair accuracy for Delphi groups, partly in line with H2. However, hidden agendas do not affect the accuracy of FTF groups, contradicting H3.

For the primary analysis, I quantify accuracy by juxtaposing group judgments to objectively true probabilities.[12] I focus on absolute error (AE)[13] as a simple, and illustrative corresponding metric:

$$AE_{g,r} = |p_{g,r} - p_r^*| \tag{1}$$

where $p_{g,r}$ is the group judgment, i.e., the estimated probability of group g in round r with $r = 1, ..., 10$, and p_r^* the corresponding true probability. Absolute errors are bound by 0, most accurate, and 1, least accurate.[14]

Absolute errors in each experimental condition are summarized in Figure 2 and lead to the three main results below. For comparisons between conditions, I report p-values of two-sided, non-parametric, adapted Wilcoxon Rank Sum tests that control for clustering of observations at the group level (Rosner et al., 2003; Jiang et al., 2020).[15]

First, I focus on the baseline scenario of situations without hidden agendas.

---

[12]In the setting of the experiment, the objectively true probability is the probability of reaching the target, which is coded into the computer program that generates the movement paths of the ladybird. An alternative approach is comparing a group judgment to the outcomes drawn according to the underlying actual probability. In the setting of the experiment, the actual outcome is the outcome reached when the program is executed for the next time, i.e., the ladybird did (not) reach the target in this particular execution of the program. This alternative, however, is a more noisy measure of accuracy, especially for relatively small numbers of distinct judgments.

[13]The reported main results 1-3 are robust towards considering squared errors and Brier scores (preregistered), as alternative metrics of accuracy (see Appendix B).

[14]An absolute error of 1, however, is only possible in two specific cases where the estimate was 0 (1) and the true probability was 1 (0), respectively. A realistic negative benchmark absolute error is 0.21, the average absolute error obtained if estimates were always 0.5, and the true probabilities are as in the experiment.

[15]Results 2 and 3 are robust to an even more conservative scenario, where the average absolute error over all ten judgments per group is considered as only independent observation per group as reported in Appendix B.

Figure 2: Absolute errors of group judgments across conditions of the estimation experiment
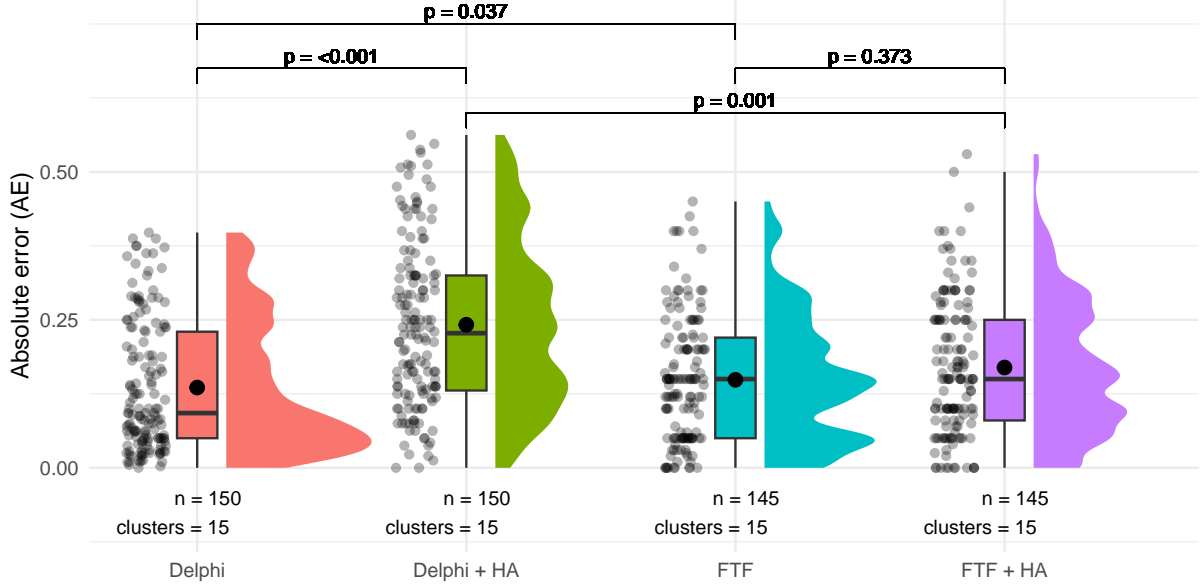
Result 1: *In line with H1, without hidden agendas, the absolute errors of judgments from structured Delphi interaction are significantly smaller (p = 0.037) than from unstructured face-to-face interaction.*

Turning to situations with hidden agendas, FTF however performs quite well.

Result 2: *Introducing hidden agendas leads to a significant increase (p < 0.001) in absolute errors of judgments from Delphi interaction but in contrast to **H2**, not for judgments from unstructured face-to-face interaction (p = 0.373).*

Result 2 also precludes that the negative effect of hidden agendas on accuracy is relatively smaller for Delphi groups. The negative effect of hidden agendas appears to be smaller in FTF than in Delphi interaction. This poses the new question: Which interaction format ultimately yields more accurate results in situations with hidden agendas?

Result 3: *With hidden agendas, the absolute errors of judgments from unstructured face-to-face interaction are significantly smaller (p = 0.001) than from structured Delphi*

Figure 3: Absolute errors of probability judgments from FTF groups against best and worst benchmarks
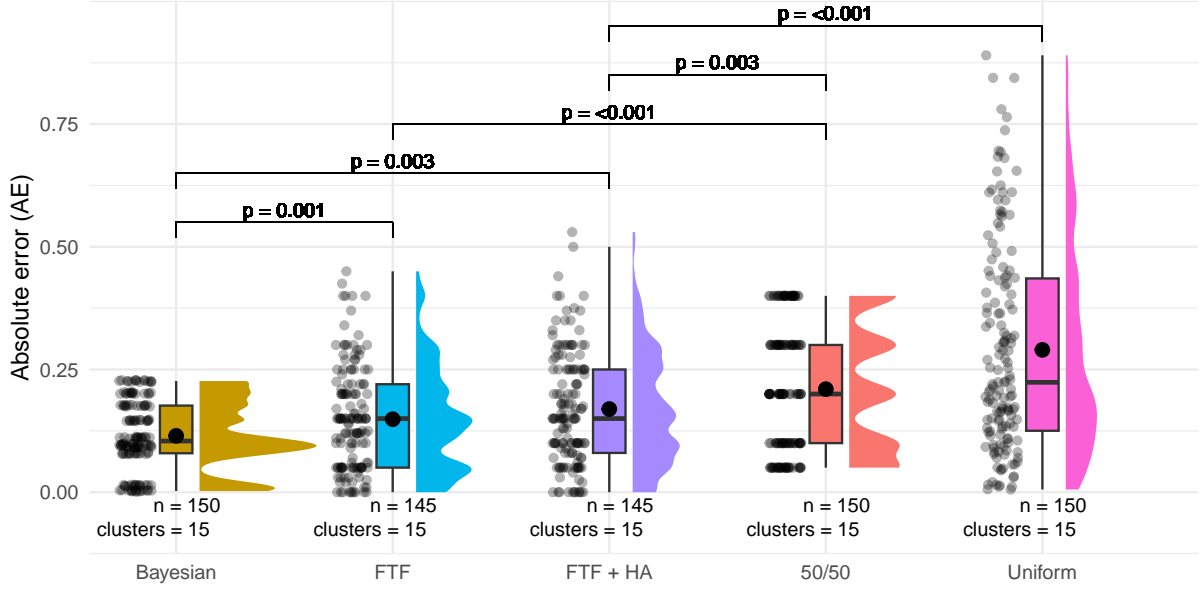
Notes: P-values from two-sided Wilcoxon rank sum tests (Rosner et al., 2003; Jiang et al., 2020). Bar in boxplot = median; dot in boxplot = mean. Observations of Bayesian, 50/50, and Uniform are simulated according to the underlying benchmark heuristic.

interaction.

4.2.1 Accuracy benchmarks

To set Results 1 to 3 into perspective, I compare absolute errors of actual group judgments to theoretical benchmarks in Figure 3 and Figure 4, respectively. On the positive side, I consider the perfect Bayesian benchmark (Bayesian). On the negative side, I consider first the naïve heuristic of always stating a "50/50" likelihood (50/50) and second reporting a random likelihood drawn from a uniform distribution (Uniform). Without hidden agendas, Delphi groups are statistically indistinguishable from the Bayesian benchmark on the positive side. FTF groups are in between positive and negative accuracy benchmarks. With hidden agendas, FTF remains more accurate than 50/50 and uniform, while Delphi is less accurate than 50/50.

The positive, Bayesian benchmark (Bayesian) considers groups as if they had a uniform, i.e., uninformative prior, on the probabilities that a step in the movement path of the judgment task takes the values -1, 0, and 1, respectively. In other words, any

19

Figure 4: Absolute errors of probability judgments from Delphi groups against best and worst benchmarks
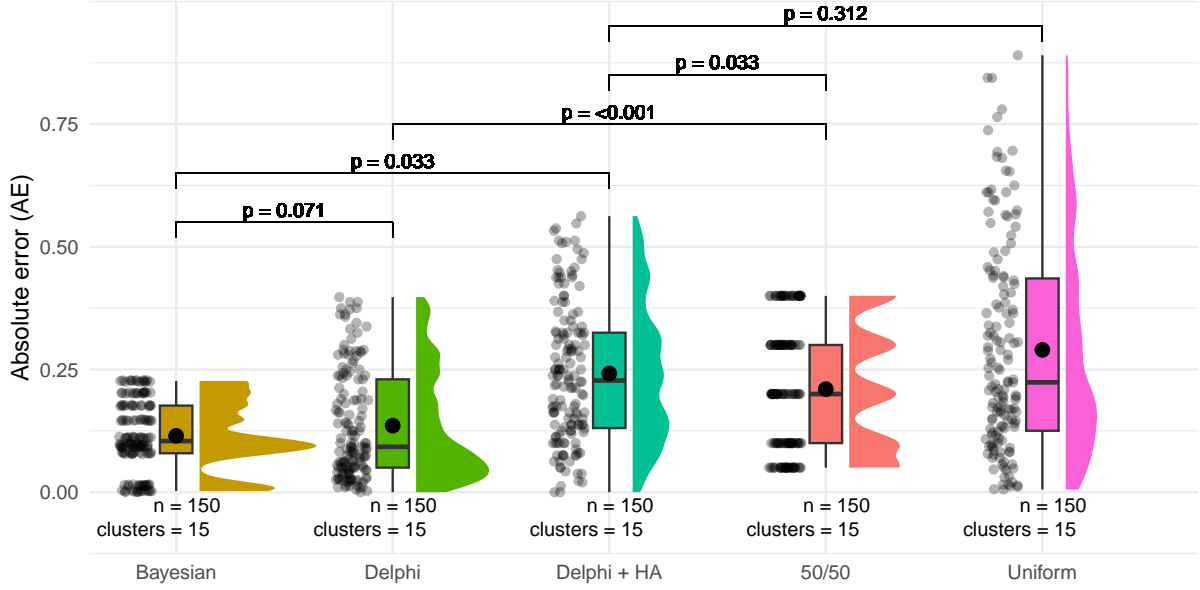
Notes: P-values from two-sided Wilcoxon rank sum tests (Rosner et al., 2003; Jiang et al., 2020). Bar in boxplot = median; dot in boxplot = mean. Observations of Bayesian, 50/50, and Uniform are simulated according to the underlying benchmark heuristic.

combination of $\mathbb{P}(X_t = -1) + \mathbb{P}(X_t = 0) + \mathbb{P}(X_t = 1) \equiv 1$ is deemed a priori equally likely. Groups updated this prior based on the observed movement path and then calculated the probability of reaching the target $\mathbb{P}(\sum_{t=1}^{10} X_t > 0)$ based on their posterior.

Result 1a: *Without hidden agendas, the absolute errors of judgments from Delphi interaction are statistically indistinguishable ($p = 0.071$) from perfect Bayesian groups, but FTF groups are significantly less ($p = 0.001$) accurate than the Bayesian benchmark.*

As the naïve 50/50 benchmark, I consider groups as if they always reported a 50% likelihood of reaching the target, irrespective of the task. Uniform constitutes an even stronger negative benchmark. It considers groups as if they consistently reported a random draw from a uniform distribution between 0 and 1, i.e., the infamous dart-throwing chimpanzee (Tetlock, 2005).

Result 1b: *Without hidden agendas, both FTF ($p < 0.001$) and Delphi ($p < 0.001$) groups are more accurate than 50/50 and Uniform, respectively.*

Result 2a: *With hidden agendas, Delphi groups are less accurate ($p = 0.033$) than groups who would always report 50% as their group judgment but remain significantly more accurate ($p = 0.006$) than the Uniform benchmark. In contrast, FTF groups remain more accurate ($p = 0.003$) than 50/50 and Uniform, even with hidden agendas.*

To summarize the findings on the accuracy, I borrow the generalist vs. specialist distinction from the field of ecology von Meijenfeldt et al. (2023). The Delphi technique appears to be a specialist that excels in a small niche or habitat (situations without hidden agendas), where it is perfectly adapted to environmental conditions. In contrast, FTF-based judgments are comparable to generalists. They never reach the same levels of accuracy as the specialist Delphi but appear way more robust and resilient to (unfavorable) changes in environmental conditions (hidden agendas).

4.2.2 Accuracy and hidden agenda achievement over time

Previously presented results focus on the accuracy of group judgments, irrespective of whether the judgment results from the very first interaction of that group or from their interactions in later rounds. In this section, I zoom in on the time dimension to analyze whether accuracy changes over time. Note that participants did not receive any feedback between rounds. Nevertheless, a potential explanation for improving accuracy is that groups may learn to interact better and enhance their understanding of the task. In the same way, in groups with hidden agendas, those group members with hidden agendas may learn how to achieve their hidden agenda over time, thus impairing accuracy. On the group level, I find no evidence for significant changes in the accuracy of judgments over time. Moreover, there is no time trend in how well group members with hidden agendas achieve their hidden agendas.

Separating group judgments for the consecutive rounds 1,2, ..., and 10 of the judgment task, I visualize accuracy and hidden agenda achievement over time in Figure 5. Based on a general overview, in most rounds, the distribution of absolute errors tends towards higher absolute errors for Delphi groups with hidden agendas, towards lower absolute errors for Delphi groups, and the distribution of absolute errors for FTF groups with and without hidden agendas falls in between. This aligns with the general results on accuracy

Figure 5: Accuracy of group judgments and hidden agenda achievement over time
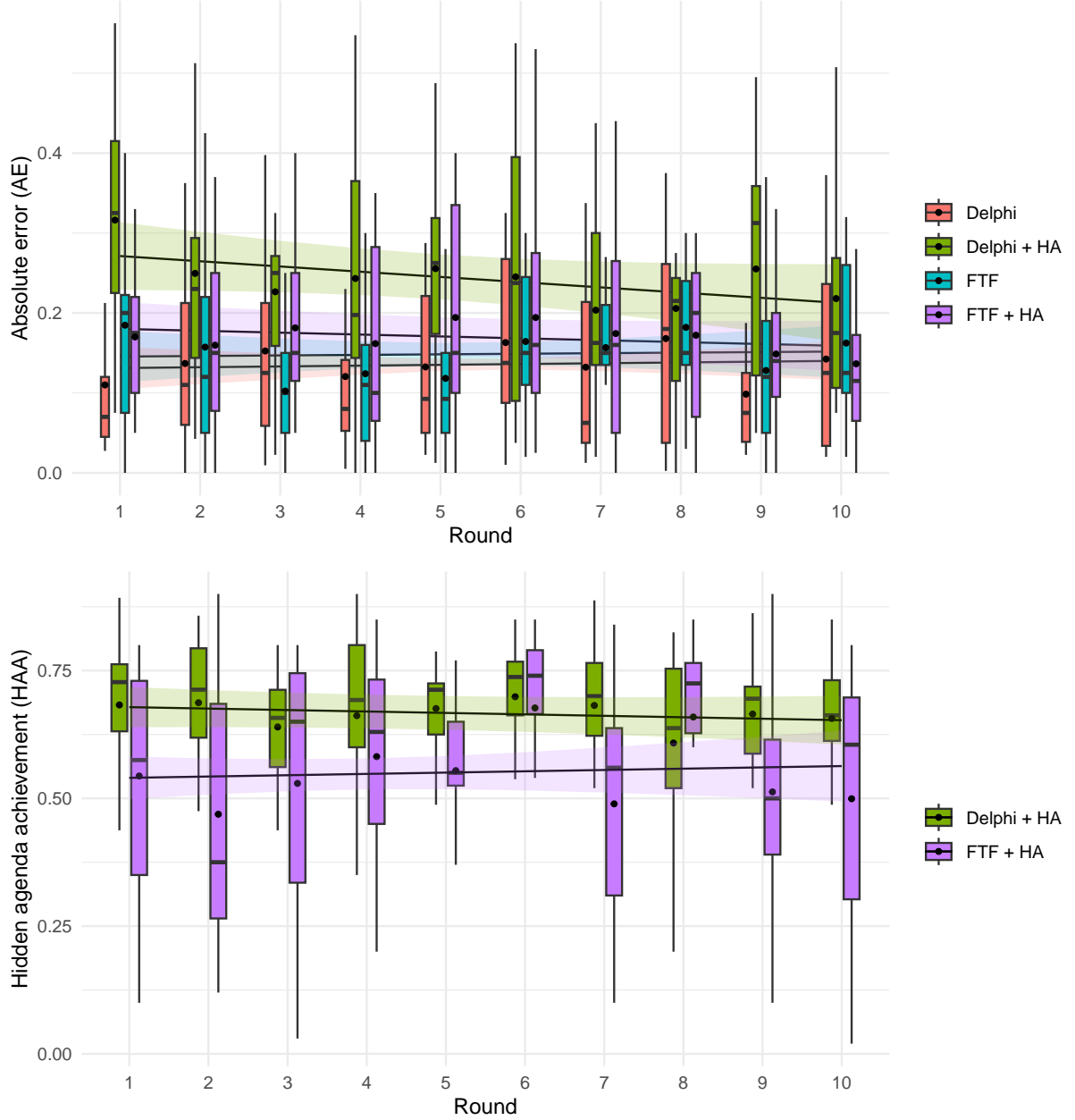
Notes: Whiskers in the boxplots span observations larger than $Q1 - 1.5 * IQR$ and smaller than $Q3 + 1.5 * IQR$. Regressions are OLS with standard errors clustered at the group level.

aggregated over all rounds as depicted earlier in Figure 2. To identify potential trends over time, I regress the absolute error as the dependent variables on the round as the independent variable. This reveals no statistically significant association between these variables. The linear regression lines accompanied by their 95%-confidence intervals are included in Figure 5. None of them has a slope that is significantly different from zero.

To quantify the degree to which group members with hidden agendas achieve their hidden agendas, I define the hidden agenda achievement rate HAA, dependent on the hidden agenda HA and the group judgment p of group g in round r as follows:

$$HAA_{g,r} = 1 - |HA_{g,r} - p_{g,r}|, \tag{2}$$

In this way, $HAA = 1$ corresponds to complete, i.e., 100%, achievement of the hidden agenda, and $HAA = 0$, indicates 0% achievement of the hidden agenda. $HA = 1$ if the hidden agenda was to drive group judgments up, and $HA = 0$ if the hidden agenda was to drive group judgments down. In most rounds, the distribution of hidden agenda achievement tends towards higher achievement rates for Delphi groups with hidden agendas and lower achievement rates for FTF groups with hidden agendas. This corroborates the general picture that FTF groups with hidden agendas are more accurate than Delphi groups with hidden agendas when aggregated over all rounds, as shown in Figure 2. Just as for accuracy, I find no statistically significant association between hidden agenda achievement and rounds.

I can extend the analysis to the individual level for Delphi groups. In any Delphi interaction, each group member provided a second individual judgment after observing their group members' first individual judgments and reasonings. The group judgment was then calculated as the mean of all second individual judgments. As previously done for the group estimates, I analyze whether these second individual judgments change accuracy over time and whether they converge more or less in the direction of hidden agendas. A priori, effects in multiple directions seem plausible. Group members with hidden agendas may, over time, distort their individual judgments more and more into the direction of the hidden agenda, as they care less about losing credibility in the decreasing number of future interactions with their group. Alternatively, they might distort less over time as they become more cautious not to behave in an overly suspicious way repeatedly. Group members without hidden agendas may, over time, update their suspicions about who has a hidden agenda and form beliefs about the direction of the hidden agenda. Based on these, they may try to counteract by distorting their own individual judgments in the opposing direction. I find no evidence for either of the two trends.

In Figure 6, I depict individual-level absolute errors and hidden agenda achievement over time. Overall, the distribution of absolute errors tends towards higher absolute

Figure 6: Accuracy and hidden agenda achievement in second individual judgments in Delphi groups over time
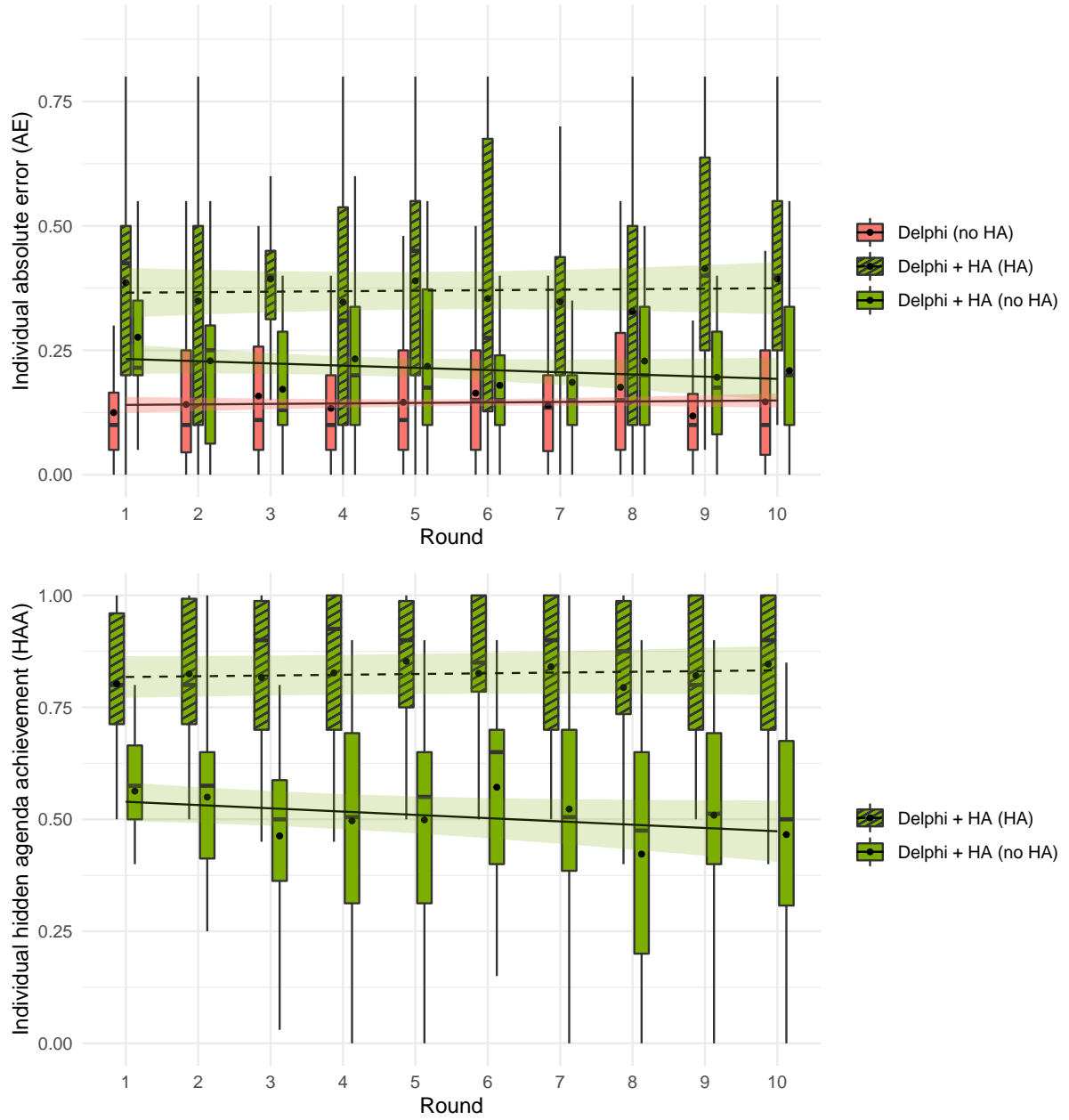
Notes: In Delphi groups without hidden agendas (Delphi), all four individual group members have no hidden agenda (no HA). In Delphi groups with hidden agendas (Delphi + HA), two individual group members have a hidden agenda (HA), and the remaining two individual group members have no hidden agenda (no HA).

Whiskers in the boxplots span observations larger than $Q1 - 1.5 * IQR$ and smaller than $Q3 + 1.5 * IQR$. Regressions are OLS with standard errors clustered at the individual level.

errors for group members with hidden agendas in Delphi groups with hidden agendas, towards lower absolute errors for group members in Delphi groups without hidden agendas, and group members without hidden agendas in Delphi groups with hidden agendas are situated in between. This suggests that group members with hidden agendas do not only negatively influence the accuracy of Delphi group judgments but also the second individual judgments of members without hidden agendas in those groups. Further, in Delphi groups with hidden agendas, individual judgments of group members with hidden agendas are plausibly shifted more in the direction of the hidden agenda than those of group members without hidden agendas. Focusing on the time dimension, I regress the individual-level absolute error and hidden agenda achievement rate on rounds. This reveals that individual accuracy and hidden agenda achievement are stable over rounds. All regression lines have a slope that is statistically indistinguishable from zero.

4.3 Evaluating trust in estimates

I measure the causal effect of hidden agendas and interaction formats on trust in group judgments by comparing confidence intervals stated in the decision experiment around group judgments from the estimation experiment. Based on the decision experiment, FTF groups always appear more trustworthy than Delphi groups, contradicting H4. Hidden agendas impair trustworthiness, in line with H5, and the negative effect is of similar magnitude for FTF and Delphi, supporting H6.

I consider the width of confidence intervals to quantify the trustworthiness of group judgments. The narrower the confidence interval, the more trustworthy the group's judgment. Confidence interval width has a most trusting upper bound at 0, or in other words, the decision-maker is certain that the group judgment is precisely equal to the actual value. The least trusting lower bound is 100, or in other words, the group judgments do not contain any information, and the decision-maker deems all possible values from 0 to 100 equally likely to be true. Distributions of confidence interval length for judgments from each condition of the estimation experiment are summarized in Figure 7. For trustworthiness differences between conditions, I report p-values of non-parametric signed rank tests adapted for clustered paired data (Datta and Satten, 2008; Jiang et al., 2020).[16]

---

[16]The Results 4 and 5 are robust towards the more conservative test, where the average interval length over all ten evaluated group judgments from the same condition of the estimation experiment, per

Figure 7: Width of confidence intervals around group judgments from different treatment conditions in the estimation experiment
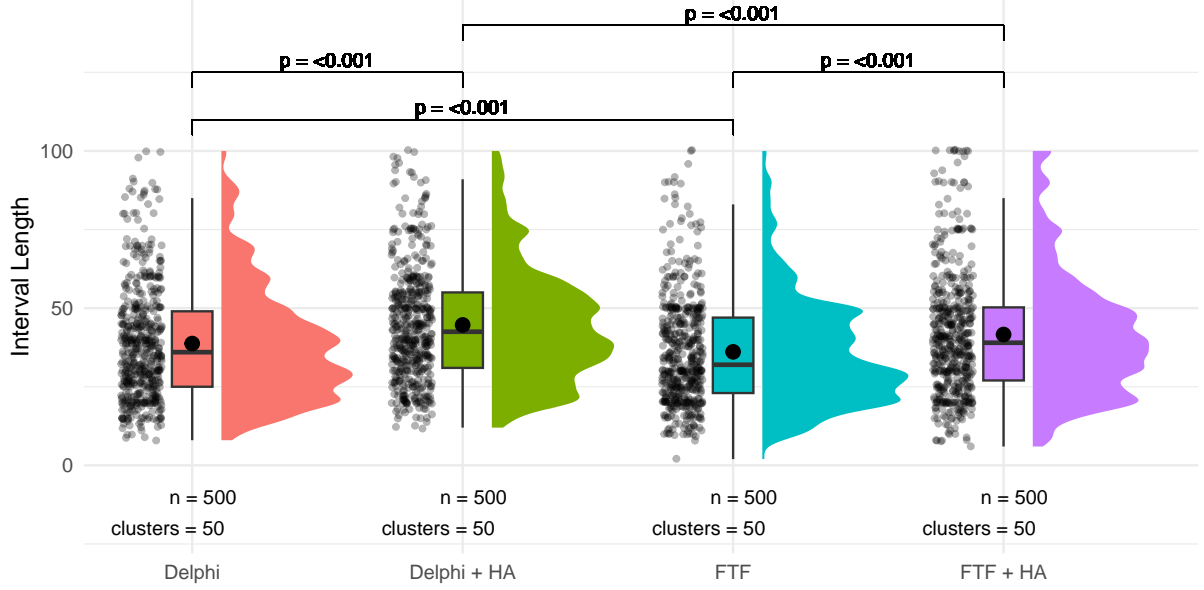
Notes: P-values from from clustered signed rank tests (Datta and Satten, 2008; Jiang et al., 2020). Bar in boxplot = median; dot in boxplot = mean.

Result 4: *In contrast to H4, FTF group judgments enjoy significantly higher ($P < 0.001$) levels of trust, i.e. narrower stated confidence intervals, than Delphi group judgments. This holds with and without hidden agendas.*

Result 5: *In line with H5, introducing hidden agendas leads to significantly lower ($P < 0.001$) levels of trust, i.e., wider confidence intervals, both for judgments from Delphi and FTF interaction.*

Zooming in on the size of the negative effect of hidden agendas on trustworthiness, I calculate the average confidence interval width for group judgments with and without hidden agendas evaluated by a specific decision-maker. Figure 8 depicts the average interval width per decision-maker, where each decision-maker is represented by two points, one for their evaluation of FTF groups and one for Delphi groups, each with (vertical axis) and without hidden agendas (horizontal axis). The distance to the 45-degree line measures the difference in trustworthiness with and without hidden agendas. As such,

---

evaluating decision-maker is considered as only independent observation (compare Appendix B)

Figure 8: Average width of confidence intervals per decision-maker around judgments of estimation experiment groups with and without hidden agendas
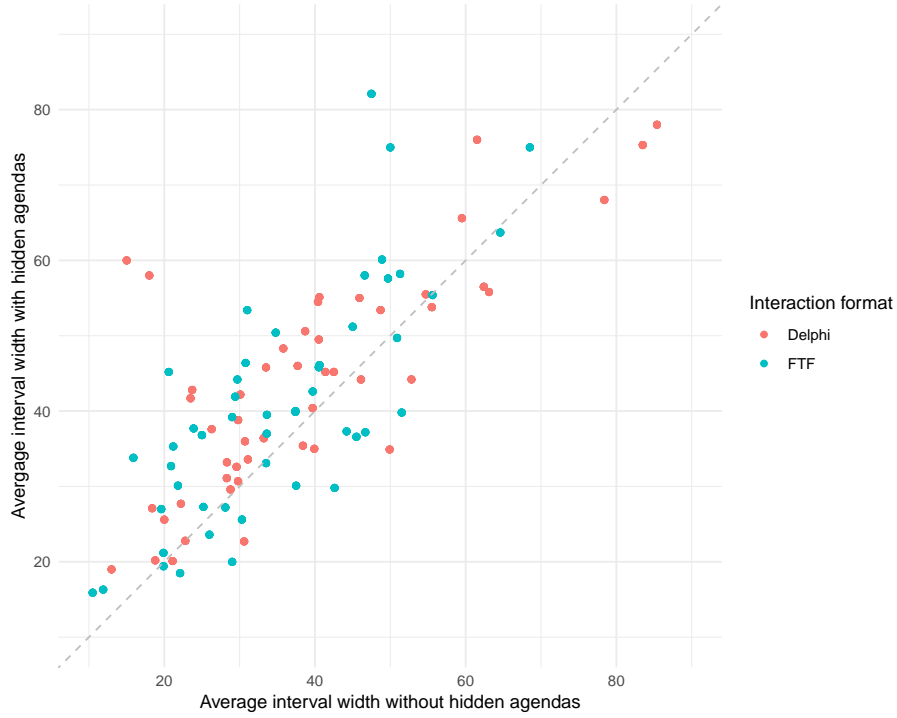
points above the 45-degree line correspond to less trustworthiness with hidden agendas than without hidden agendas. In line with Result 5, we see that more points lie above the 45-degree lines, and many are further away from the line than points below the 45-degree line. Further, suppose the impact of hidden agendas on trustworthiness is similar for FTF and Delphi. In that case, we should see no clusters of points representing FTF and Delphi group judgments, e.g., Delphi points generally lie further above the 45-degree line than FTF points. In line with Result 6, points are scattered along the 45-degree line without clearly visible clusters.

Result 6: *In line with H6, the size of the negative effect of hidden agendas on trust in judgments from FTF groups and Delphi groups is statistically not distinguishable (p = 0.6328).*

Combining results on accuracy and trustworthiness, the ultimate question is: Does trustworthiness align with accuracy, i.e., do decision-makers correctly put more trust in the more accurate judgments and vice versa?

Result 7: *Trustworthiness is aligned with accuracy in settings with hidden agendas, the objectively more accurate judgments from FTF interaction are trusted more. However, trustworthiness is misaligned with accuracy in settings without hidden agendas, the objectively less accurate judgments from FTF interaction are trusted more.*

Furthermore, the changes in trustworthiness are misaligned with changes in objective accuracy. While Delphi group judgments become relatively less accurate than FTF group judgments through the introduction of hidden agendas, trust in Delphi group judgments decreases just the same as trust in FTF group judgments.

4.4   BIN-model

To formally disentangle the mechanisms behind accuracy differences between group interaction formats in settings with and without hidden agendas, I use the BIN-model of forecasting (Satopää et al., 2021). This model distinguishes three determinants of the accuracy of probabilistic judgments or estimates: bias, information usage, and noise (BIN) as latent, i.e., not directly observable, variables of an estimating group. Bias resembles a group's tendency to produce too high or too low estimates consistently, noise leads to non-systematic variability in estimates unrelated to the event of interest, and finally, information usage characterizes the group judgments' variability, which is correlated with the event of interest. Bias and noise are detrimental, while information usage is beneficial to accuracy. Given the modeling framework, all three latent variables can be structurally estimated based on group judgments and compared across conditions of the estimation experiment. Below, I briefly summarize the BIN-model framework. Further, I outline adaptions needed to accommodate the present setting of group judgments instead of individual forecasters in the original setting. For a more detailed discussion of the model, I refer the reader to the original paper (Satopää et al., 2021).

Let the binary event of interest of estimation be denoted as $Y \in \{0, 1\}$, i.e. (not) reaching the target. Specifically, $Y = 1$ if $\sum_{t=1}^{10} X_t > 0$, i.e. the target is reached, and $Y = 0$ otherwise. Further, the model builds on two pillars: objective signals Z^* about Y and human estimates of the likelihood of Y based on the subjective interpretation of these signals Z. The outcome Y is considered to be determined by the entirety of all past

and future objective signals, modeled as the continuous, normally distributed variable $Z^* \sim \mathcal{N}(\mathbb{E}[Z^*], Var(Z^*))$. In particular, the event is assumed to happen, i.e., the target area is reached, if the sum of all these signals is positive $Y = \mathbf{1}(Z^* > 0)$ with the indicator function $\mathbf{1}(E)$ equal 1 if E is true, and 0 otherwise. Intuitively, this can be interpreted as an accumulation of evidence in favor of the event happening. The model fixes the mean $\mathbb{E}[Z^*] = \mu^*$ and $Var(Z^*) = 1$, such that $\mathbf{P}(Z^* > 0) = p^*$, i.e. the expected frequency of the event happening based on Z^* is aligned with the true base rate $p^* \in (0, 1)$.

Groups in a specific (experimental) condition i judge the likelihood of Y dependent on their interpretation of signals Z_i. Just as Z^*, Z_i is modeled as a normally distributed variable subject to three latent variables of group accuracy: bias, information usage usage, and noise (BIN). Intuitively, the closer the subjective Z_i is to the objective Z^*, the more accurate the group's judgments. Inaccuracy can take the form of bias as the difference in the means of Z^* and Z_i, or noise as the variability of Z^* that is uncorrelated with Z_i. Accuracy-enhancing information usage is described by the covariance of Z^* and Z_i. More formally, Z_i and Z^* follow the multivariate normal distribution:

$$\begin{pmatrix} Z^* \\ Z_i \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \mu^* \\ \mu^* + \mu_i \end{pmatrix} \begin{pmatrix} 1 & \gamma_i \\ \gamma_i & \gamma_i + \delta_i \end{pmatrix} \right) \tag{3}$$

where the outcome and the latent variables are defined as:

$$Outcome: \quad Y = \mathbf{1}(Z^* > 0) \tag{4}$$

$$Bias: \quad \mu_i = \mathbb{E}[Z_i] - \mathbb{E}[Z^*] \tag{5}$$

$$Information: \quad \gamma_i = Cov(Z_i, Z^*) \tag{6}$$

$$Noise: \quad \delta_i = Var(Z_i) - Cov(Z_i, Z^*), \tag{7}$$

where the subscript i denotes a specific condition in the estimation experiment (FTF or Delphi interaction paired with (no) hidden agendas). For interpretation, perfectly accurate groups would be unbiased with $\mu_i = 0$, noise-free with $\delta_i = 0$, and exhibit $\gamma_i = 1 = Var(Z^*)$. Bias increases the further μ_i is from 0, noise increases the larger δ_i, and information usage decreases the further γ_i below 1.

Translating signals into probability judgments, the model considers groups as reporting their rational Bayesian belief of Y given Z_i:

$$\mathbb{E}[Y|Z_i] = \mathbb{P}[Z^* > 0|Z_i] \tag{8}$$

Groups are treated as if they report a judgment after they try their best to eliminate bias and noise. They are unaware of any remaining bias and noise in their interpretation of signals.[17] Given this bounded rationality assumption, group judgments are given by:[18]

$$\mathbb{P}[Z^* > 0|Z_i] = \Phi\left(\frac{Z_i}{\sqrt{1 - \gamma_i}}\right), \tag{9}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

I use this framework to compare bias, information usage, and noise of groups across conditions of the estimation experiment. In this way, I can quantify the causal effect of the interaction format and hidden agendas on these three latent variables of accuracy. To this end, the estimation process of groups in the same condition is assumed to be subject to the same bias, information usage, and noise. In other words, the estimation does not aim at the individual group level but treats any given group in a specific (experimental) condition as representative, i.e., as interchangeable with any other group in this condition. Further, outcomes and predictions across the ten rounds of the judgment tasks are assumed to be independent. Consequently, an estimate of, e.g., $\mathbb{P}[Y = 1] = 0.3$ in the first round of the judgment task says nothing about the estimate in the next round, except for the group's latent bias, information usage, and noise. As such, parameter estimates for bias, information usage, and noise can be seen as averages within a specific condition, e.g., the average bias of FTF groups with hidden agendas across the ten judgment tasks in the estimation experiment.

4.4.1 Estimates of bias, information usage, and noise

Following the econometric procedures of Satopää et al. (2021), I estimate the parameters in the framework of the BIN-model using Bayesian statistics. To quantify the causal

---

[17]This seems reasonable, as groups are incentivized through a proper scoring rule to produce the most accurate estimate possible.

[18]See Satopää et al. (2021) for the derivation of the conditional probability and a discussion of this assumption.

effects of interaction formats and hidden agendas on bias, information usage, and noise, I consider four binary comparisons of conditions in the estimation experiment. On the one hand, two comparisons juxtapose interaction formats within the same incentive scheme: *FTF vs. Delphi* and *FTF with hidden agendas (FTF+HA) vs. Delphi with hidden agendas (Delphi+HA)*. On the other hand, two comparisons juxtapose incentive schemes within the same interaction format: *FTF vs. FTF+HA* and *Delphi vs. Delphi+HA*. For each binary comparison, I estimate posterior parameter distributions based on a uniform prior. In other words, a prior, I make the conservative assumption that all parameter configurations in the BIN-model are equally likely across the two compared conditions.[19] To derive posterior distributions, parameter estimates are updated according to Bayes rule, taking observed group judgments, i.e., the data, into account. I modify the BINtools package (Satopää et al., 2022) to derive these parameter estimates by implementing Markov Chain Monte Carlo methods in R and Stan.[20] Table 2 presents estimates of the BIN-model parameters and posterior probabilities. Further, details on the estimation are outlined in Appendix C.1.

To structure results, I focus on two guiding questions consecutively: First, what is the likelihood of changes in accuracy that can be attributed to the mechanisms: bias, information usage, and noise? Second, what is the expected magnitude of changes in accuracy that can be attributed to each mechanism?

Regarding the likelihood of changes, noise appears as the most likely mechanism driving accuracy differences between interaction formats. In contrast, differences between incentive schemes are most likely driven by the use of valuable information and bias. Estimates of the likelihoods are presented in Table 2 in the section on posterior inferences. Posterior probabilities are the Bayesian equivalent of p-values in frequentist statistics. The larger the posterior probability, the stronger the evidence in favor of the hypothesis and vice versa.

---

[19]This prior is relatively uninformative, and hence parameter estimates will largely be influenced by observed, actual group judgments.

[20]Prior to the estimation, I aligned all group judgments such that any hidden agenda points in the direction of driving group judgments up. This ensures that the estimates of bias can be consistently interpreted in relation to the direction of the hidden agenda. Positive estimates of bias indicate that judgments are systematically distorted in the direction of hidden agendas and vice versa.

Table 2: Bayesian estimates of posterior inferences and BIN model parameters

Summary Statistic	Interaction	Hidden Agendas		Interaction
	FTF vs. Delphi	FTF vs. FTF+HA	Delphi vs. Delphi+HA	FTF+HA vs. Delphi+HA
Posterior inferences				
Less bias in treatment: $\mathbb{P}(|\mu_1| < |\mu_0|)$	0.737	0.506	0.187	0.149
Less noise in treatment: $\mathbb{P}(\delta_1 < \delta_0)$	0.997	0.598	0.75	0.993
More info in treatment: $\mathbb{P}(\gamma_0 < \gamma_1)$	0.221	0.288	0.177	0.158
Parameter estimates (with 95% CI)				
Outcome mean: μ^*	0.00 (-0.24; 0.23)	-0.01 (-0.24; 0.23)	0.00 (-0.24; 0.23)	0.00 (-0.23; 0.23)
Bias (control): μ_0	-0.10 (-0.72; 0.47)	-0.11 (-0.76; 0.47)	-0.09 (-0.53; 0.47)	0.11 (-0.48; 0.47)
Bias (treatment): μ_1	-0.09 (-0.51; 0.31)	0.11 (-0.51; 0.31)	0.44 (0.09; 0.31)	0.43 (0.09; 0.31)
Diff in Bias: $|\mu_0| - |\mu_1|$	0.06 (-0.10; 0.29)	0.01 (-0.30; 0.29)	-0.26 (-0.59; 0.29)	-0.18 (-0.46; 0.29)
Information (control): γ_0	0.14 (0.01; 0.40)	0.15 (0.01; 0.40)	0.12 (0.01; 0.40)	0.15 (0.01; 0.40)
Information (treatment): γ_1	0.10 (0.01; 0.27)	0.12 (0.01; 0.27)	0.07 (0.00; 0.27)	0.07 (0.00; 0.27)
Diff in information: $\gamma_0 - \gamma_1$	0.05 (-0.06; 0.17)	0.03 (-0.08; 0.17)	0.05 (-0.06; 0.17)	0.08 (-0.06; 0.17)
Noise (control): δ_0	0.94 (0.19; 3.21)	0.96 (0.15; 3.21)	0.29 (0.01; 3.21)	0.83 (0.12; 3.21)
Noise (treatment): δ_1	0.31 (0.02; 1.24)	0.90 (0.15; 1.24)	0.18 (0.02; 1.24)	0.17 (0.01; 1.24)
Diff in noise: $\delta_0 - \delta_1$	0.63 (0.12; 2.10)	0.06 (-0.34; 2.10)	0.12 (-0.10; 2.10)	0.66 (0.05; 2.10)

Control: condition named on top, Treatment: condition named below CI: credible intervals

Comparing FTF to Delphi group judgments, Delphi group judgments exhibit less noise with very high probabilities ($\mathbb{P} > 0.99$) in situations with and without hidden agendas. A potential reason might be that Delphi group judgments constitute an average of four individual judgments. In contrast, in FTF interaction, the group judgment constitutes a consensus judgment among the four group members.

Result 8: *Noise is the most likely mechanism driving accuracy differences between interaction formats. Delphi group judgments are likely less noisy than FTF group judgments.*

Comparing situations without and with hidden agendas within the same interaction format, less usage of valuable information is the most likely mechanism contributing to a decline in accuracy for FTF ($\mathbb{P} = 0.712$) and even more so for Delphi groups ($\mathbb{P} = 0.823$).

For Delphi groups with hidden agendas, an increase in bias is about as likely ($\mathbb{P} = 0.813$) as a decrease in the usage of valuable information. This is not mirrored in FTF groups, which are almost equally likely to exhibit less or more bias with hidden agendas compared to situations without. This points towards robustness against bias as differentiating mechanisms that may explain why Delphi groups do and FTF group judgments do not become significantly less accurate in situations with hidden agendas. Nevertheless, posterior probabilities for the effect of hidden agendas are far from any usual significance level used when interpreting p-values in frequentist settings. Statements on the likelihood of information usage and bias driving accuracy differences between situations with and without hidden agendas should thus be interpreted cautiously.

Result 9: *Less use of valuable information is the most likely mechanism for decreasing accuracy with hidden agendas irrespective of the interaction format.*

Result 10: *Increased bias likely decreases accuracy for Delphi group judgments in situations with hidden agendas. FTF group judgments appear robust against this effect.*

In terms of magnitudes of changes, noise emerges as the only mechanism with a clear directional effect on accuracy differences between interaction formats. Estimates of the effect sizes of bias, information usage, and noise on accuracy are presented in the lower part of Table 2. Of particular interest are estimates for the difference in bias ($|\mu_1| < |\mu_0|$), information usage ($\gamma_0 - \gamma_1$), and noise ($\delta_1 - \delta_0$).

Result 10: *Noise is the only mechanism to which a clear directional effect can be attributed. FTF group judgments exhibit more noise than Delphi group judgments.*

Comparing FTF to Delphi group judgments, Delphi group judgments exhibit less noise in situations without ($\delta_0 - \delta_1 = 0.63$) and with hidden agendas ($\delta_0 - \delta_1 = 0.66$). The respective 95% credible intervals around the estimates of $\delta_0 - \delta_1$ do not include zero. Effects of information usage and bias are estimated with 95% credible intervals spanning zero. Thus, I cannot identify a clear directional effect of these mechanisms in the BIN-model framework.

Comparing situations without and with hidden agendas within the same interaction format, neither bias, information usage, nor noise can be identified with a clear directional effect. The respective 95% credible intervals around the estimates span zero.

Wide credible intervals for magnitudes of expected effects can likely be attributed to the relatively small number of judgment tasks solved per group in the estimation experiment. Satopää et al. (2021) show that the width of credible intervals decreases substantially if the number of forecasts by the same person increases. This holds especially for relatively low numbers of below 50 forecasts. While Satopää et al. (2021) use the Good Judgment Project data, which comprises between 87 and 191 forecasts per person, the estimation experiment yields only ten distinct judgments per group.[21]

4.5 Communication Patterns

The BIN-model estimates are a good starting point to shed light on potential mechanisms underlying accuracy differences across conditions of the estimation experiment. Yet, essential differences might not be uncovered by structural estimations alone. To complement the BIN-model estimations, I analyze the degree and truthfulness of information sharing during group interaction. To this end, the communication of FTF and Delphi interactions with and without hidden agendas are transcribed and coded.[22]. On a high level, the communication protocols reveal that hidden agendas cause less shared, truthful information in Delphi groups but not in FTF groups.

---

[21] The number of 10 distinct judgments per group ensured implementability in a controlled lab experiment. This has clear advantages for analyzing the causal effects of group interaction formats and incentive schemes on accuracy and trust. Conversely, experimental control has been traded off against a higher number of observations, which might have allowed more precise estimates of mechanisms driving accuracy differences in the BIN-model framework.

[22] A detailed outline of the transcription and coding procedure can be found in Appendix D

Figure 9: Revelation of information across conditions of the estimation experiment
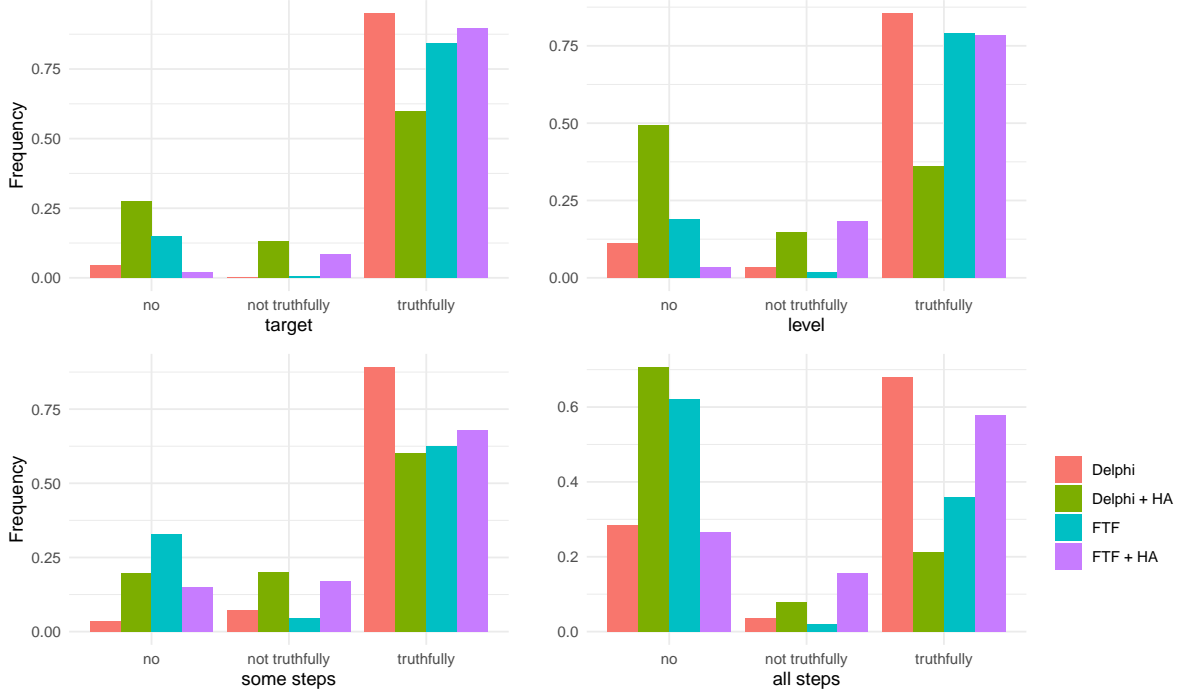
The probabilistic judgment task of the estimation experiment is generally designed such that the more information participants, i.e., group members, share with their group, the more accurately the group can solve the task. In increasing level of detail, group members can reveal whether their individual movement path reached the target (target), the final level of their movement path after ten steps (level), one or more individual steps of their movement path (some steps), and most comprehensively all ten steps of their movement path (all steps). Furthermore, for each level of detail, group members may state their information truthfully or not truthfully. Withholding or not stating information truthfully may be a strategy to drive group estimates up or down following one's hidden agenda.

In Figure 9, I summarize the frequencies of not stated (no), not truthfully stated (not truthfully), and truthfully stated (truthfully) information at the previously described four levels of detail. I estimate corresponding linear probability models to identify the significance of differences in frequencies. Estimation results are presented in Appendix D.

Focusing first on whether information is shared at all, it becomes visible that in Delphi interactions with hidden agendas, significantly more group members choose not to reveal information when compared to Delphi interactions without hidden agendas. This holds

at all levels of detail. Further, with hidden agendas, the frequency of shared non-truthful information is significantly higher, and the frequency of shared truthful information is significantly lower. Again, this holds at all levels of detail. This observation is intuitively aligned with the lower accuracy of Delphi groups with hidden agendas. Not only do hidden agendas lead to less information sharing, but they also decrease the truthfulness of shared information.

We see a vastly different pattern comparing FTF groups with and without hidden agendas. With hidden agendas, significantly more group members choose to reveal information. This holds at all levels of detail. Just as for Delphi, in FTF groups with hidden agendas, the frequency of shared non-truthful information is significantly higher on all levels of details. However, unlike Delphi, the frequency of shared truthful information is significantly higher for the target, some steps, and all steps. This observation suggests an intuitive explanation for the robust accuracy of FTF groups despite hidden agendas. While hidden agendas lead to more sharing of untruthful information, at the same time, the sharing of truthful information increases and potentially compensates the negative effects of untruthful information sharing.

5 Conclusion and discussion

In this paper, I study whether commonly used face-to-face meetings and the scientifically supported Delphi technique are suitable interaction formats to generate and elicit accurate and trustworthy group judgments. In particular, I consider situations where some group members have a hidden agenda, i.e., incentives to manipulate. Through two complementary experiments, I provide evidence supporting the widespread use of FTF meetings in real-world institutional decision-making. Hidden agendas are a potential threat to the accuracy and trustworthiness of group judgments. However, FTF emerges as capable of mitigating this threat. FTF interaction appears to be a resilient generalist capable of generating accurate and trusted group judgments even under adverse conditions with hidden agendas. In contrast, the Delphi technique appears to be a specialist adapted to one niche, capable of peak performance and outperforming FTF, but only without hidden agendas and only in terms of accuracy.

Using structural estimations in the Bayesian BIN-model framework, I pinpoint increased bias as an accuracy-impairing mechanism that may explain differences between

FTF and Delphi. While Delphi group judgments likely suffer from increased bias toward the direction of hidden agendas, FTF group judgments do not exhibit this effect. Robustness towards bias, thus, seems a relevant concern when choosing and designing if and how decision-informing group judgments are generated and hidden agendas might play a role. This study provides a building block of evidence from a controlled lab experiment that favors employing unstructured FTF group judgments in these situations. However, while FTF can be identified as relatively better than Delphi, it remains an open question whether unstructured FTF is the best among a wider set of interaction formats in situations with hidden agendas and whether the result holds in different contexts.

Further research may thus zoom in on identifying particular features of FTF underlying its robustness towards manipulation. This may deepen our understanding of why certain existing interaction formats work better than others, but ultimately also help to design new interaction formats to better extract accurate and trustworthy group judgments. To this end, the experimental setup of this study may serve as a test bed well suited to investigate the accuracy and trustworthiness of any group interaction format directed towards generating quantitative judgments in general. To isolate the effect of particular features of an interaction format, modifications of group interaction formats that vary by one detail at a time may be tested against each other. Furthermore, the present study investigates settings where the existence of hidden agendas is common knowledge. In institutional decision making it is however more likely, that one can only suspect the presence of hidden agendas, hence there is uncertainty about their presence. Future research could expand to such settings where hidden agendas may be present probabilistically. Beyond this, the investigation of the capabilities of group interaction to generate accurate and trustworthy judgments should also be extended to settings outside of the controlled lab environment and tested in the field. A potential stepping stone could be to investigate the presence and relevance of hidden agendas in forecasting tournaments like the Good Judgment Project (Mellers et al., 2014).

Taken together, this study emphasizes that hidden agendas and potential manipulation matter and may have a detrimental effect on the accuracy and trustworthiness of expertise. It is vital to account for their existence to alleviate their threat to institutional decision-making. The presented evidence suggests that FTF is the preferable interaction format to generate accurate and trustworthy group judgments in situations with hidden agendas.

The Delphi technique is preferable in situations without information where the accuracy of group judgments is of top priority. Practitioners involved in eliciting group judgments may directly link to these results and structure the interaction format of groups accordingly.

References

ABELER, J., D. NOSENZO, AND C. RAYMOND (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–1153. Cited on page 10.

AMJAHID, M., D. MÜLLER, Y. MUSHARBASH, H. STARK, AND F. ZIMMERMANN (2017): "An Attack is Expected," *Die Zeit*, Nr. 13/2017. Cited on page 2.

ARMSTRONG, J. (2006): "How to Make Better Forecasts and Decisions: Avoid Face-to-Face Meetings," *Foresight: The International Journal of Applied Forecasting*, 5, 3–15. Cited on page 3.

ARROW, K. J., R. FORSYTHE, M. GORHAM, R. HAHN, R. HANSON, J. O. LEDYARD, S. LEVMORE, R. LITAN, P. MILGROM, F. D. NELSON, G. R. NEUMANN, M. OTTAVIANI, T. C. SCHELLING, R. J. SHILLER, V. L. SMITH, E. SNOWBERG, C. R. SUNSTEIN, P. C. TETLOCK, P. E. TETLOCK, H. R. VARIAN, J. WOLFERS, AND E. ZITZEWITZ (2008): "The Promise of Prediction Markets," *Science*, 320, 877–878. Cited on page 2.

ATANASOV, P., P. RESCOBER, E. STONE, S. A. SWIFT, E. SERVAN-SCHREIBER, P. TETLOCK, L. UNGAR, AND B. MELLERS (2017): "Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls," *Management Science*, 63, 691–706. Cited on page 6.

BECKER, J., D. BRACKBILL, AND D. CENTOLA (2017): "Network dynamics of social influence in the wisdom of crowds," *Proc Natl Acad Sci U S A*, 114, E5070–E5076. Cited on page 9.

BECKER, J. A., D. GUILBEAULT, AND E. B. SMITH (2021): "The Crowd Classification Problem: Social Dynamics of Binary-Choice Accuracy," *Management Science*. Cited on page 14.

BELTON, I., A. MACDONALD, G. WRIGHT, AND I. HAMLIN (2019): "Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process," *Technological Forecasting and Social Change*, 147, 72–82. Cited on page 9.

Belton, I., G. Wright, A. Sissons, F. Bolger, M. M. Crawford, I. Hamlin, C. Taylor Browne Lūka, and A. Vasilichi (2021): "Delphi with feedback of rationales: How large can a Delphi group be such that participants are not overloaded, de-motivated, or disengaged?" *Technological Forecasting and Social Change*, 170. Cited on page 14.

Best, R. J. (1974): "An Experiment in Delphi Estimation in Marketing Decision Making," *Journal of Marketing Research*, 11, 448–452. Cited on page 9.

Bolger, F., A. Stranieri, G. Wright, and J. Yearwood (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, 1671–1680. Cited on page 9.

Bolger, F. and G. Wright (2011): "Improving the Delphi process: Lessons from social psychological research," *Technological Forecasting and Social Change*, 78, 1500–1513. Cited on page 9.

Bond, C. F., J. and B. M. DePaulo (2006): "Accuracy of deception judgments," *Pers Soc Psychol Rev*, 10, 214–34. Cited on page 10.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree - An Open-Source Platform for Laboratory, Online, and Field Experiments," *SSRN Electronic Journal*. Cited on page 11.

Condorcet, J.-A.-N. d. C. m. d. (1785): *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la probabilite des voix*, Library of liberal arts ; LLA 159, Indianapolis: Bobbs-Merrill, 1st ed., translated in 1976 to "Essay on the Application of Mathematics to the Theory of Decision-Making,". Cited on page 10.

Conrads, J., B. Irlenbusch, R. M. Rilke, and G. Walkowitz (2013): "Lying and team incentives," *Journal of Economic Psychology*, 34, 1–7. Cited on page 10.

Cowgill, B. and E. Zitzewitz (2015): "Corporate Prediction Markets: Evidence from Google, Ford, and Firm X," *The Review of Economic Studies*, 82, 1309–1341. Cited on page 2.

DA, Z. AND X. HUANG (2020): "Harnessing the Wisdom of Crowds," *Management Science*, 66, 1847–1867. Cited on page 9.

DAAR, A. S., H. THORSTEINSDOTTIR, D. K. MARTIN, A. C. SMITH, S. NAST, AND P. A. SINGER (2002): "Top ten biotechnologies for improving health in developing countries," *Nat Genet*, 32, 229–32. Cited on page 9.

DALKEY, N. C. (1975): "Toward a theory of group estimation," *The Delphi method: Techniques and applications*, 236–261. Cited on page 9.

DATTA, S. AND G. A. SATTEN (2008): "A signed-rank test for clustered data," *Biometrics*, 64, 501–7. Cited on pages 25 and 26.

ECKEN, P., T. GNATZY, AND H. A. VON DER GRACHT (2011): "Desirability bias in foresight: Consequences for decision quality based on Delphi results," *Technological Forecasting and Social Change*, 78, 1654–1670. Cited on page 8.

FEDDERSEN, T. AND W. PESENDORFER (1998): "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting," *American Political Science Review*, 92, 23–35. Cited on page 10.

FEHRLER, S. AND N. HUGHES (2018): "How Transparency Kills Information Aggregation: Theory and Experiment," *American Economic Journal: Microeconomics*, 10, 181–209. Cited on page 10.

FELGENHAUER, M. AND H. P. GRÜNER (2008): "Committees and Special Interests," *Journal of Public Economic Theory*, 10, 219–243. Cited on page 2.

FISCHBACHER, U. AND F. FÖLLMI-HEUSI (2013): "Lies in Disguise-an Experimental Study on Cheating," *Journal of the European Economic Association*, 11, 525–547. Cited on page 10.

FRIEDMAN, J. A. AND R. ZECKHAUSER (2014): "Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden," *Intelligence and National Security*, 30, 77–99. Cited on page 2.

FRIES, T., U. GNEEZY, A. KAJACKAITE, AND D. PARRA (2021): "Observability and lying," *Journal of Economic Behavior & Organization*, 189, 132–149. Cited on page 10.

GALTON, F. (1907): "Vox populi," *Nature*, 75. Cited on page 9.

GIMPEL, H. AND F. TESCHNER (2014): "Market-Based Collective Intelligence in Enterprise 2.0 Decision Making," . Cited on page 7.

GNEEZY, U. (2005): "Deception: The Role of Consequences," *American Economic Review*, 95. Cited on page 10.

GNEEZY, U., A. KAJACKAITE, AND J. SOBEL (2018): "Lying Aversion and the Size of the Lie," *American Economic Review*, 108, 419–453. Cited on page 10.

GOEREE, J. K. AND L. YARIV (2011): "An Experimental Study of Collective Deliberation," *Econometrica*, 79, 893–921. Cited on page 10.

GRAEFE, A. AND J. S. ARMSTRONG (2011): "Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task," *International Journal of Forecasting*, 27, 183–195. Cited on pages 3, 6, 8, and 13.

GREINER, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125. Cited on page 11.

GRIME, M. M. AND G. WRIGHT (2016): *Delphi Method*, John Wiley & Sons, Ltd, 1–6. Cited on pages 9 and 14.

GRISCOM, B. W., J. ADAMS, P. W. ELLIS, R. A. HOUGHTON, G. LOMAX, D. A. MITEVA, W. H. SCHLESINGER, D. SHOCH, J. V. SIIKAMÄKI, P. SMITH, P. WOODBURY, C. ZGANJAR, A. BLACKMAN, J. CAMPARI, R. T. CONANT, C. DELGADO, P. ELIAS, T. GOPALAKRISHNA, M. R. HAMSIK, M. HERRERO, J. KIESECKER, E. LANDIS, L. LAESTADIUS, S. M. LEAVITT, S. MINNEMEYER, S. POLASKY, P. POTAPOV, F. E. PUTZ, J. SANDERMAN, M. SILVIUS, E. WOLLENBERG, AND J. FARGIONE (2017): "Natural climate solutions," *Proceedings of the National Academy of Sciences*, 114, 11645–11650. Cited on page 9.

HANSEN, J., C. SCHMIDT, AND M. STROBEL (2004): "Manipulation in political stock markets - preconditions and evidence," *Applied Economics Letters*, 11, 459–463. Cited on pages 2 and 7.

HANSON, R. AND R. OPREA (2009): "A Manipulator Can Aid Prediction Market Accuracy," *Economica*, 76, 304–314. Cited on page 8.

HANSON, R., R. OPREA, AND D. PORTER (2006): "Information aggregation and manipulation in an experimental market," *Journal of Economic Behavior & Organization*, 60, 449–459. Cited on page 8.

HAUG, N., L. GEYRHOFER, A. LONDEI, E. DERVIC, A. DESVARS-LARRIVE, V. LORETO, B. PINIOR, S. THURNER, AND P. KLIMEK (2020): "Ranking the effectiveness of worldwide COVID-19 government interventions," *Nat Hum Behav*, 4, 1303–1312. Cited on page 2.

HERMANN, D. AND M. BRENIG (2022): "Dishonest online: A distinction between observable and unobservable lying," *Journal of Economic Psychology*, 90, 102489. Cited on page 10.

HERMANN, D. AND A. OSTERMAIER (2018): "Be Close to Me and I Will Be Honest. How Social Distance Influences Honesty," *SSRN Electronic Journal*. Cited on page 10.

HOSSAIN, T. AND R. OKUI (2013): "The Binarized Scoring Rule," *The Review of Economic Studies*, 80, 984–1001. Cited on page 50.

JABOTINSKY, H. Y. AND R. SAREL (2020): "Let it Flow: Information Exchange in Video Conferences versus Face-to-Face Meetings," *Working Paper*. Cited on page 13.

JIANG, Y., M.-L. T. LEE, X. HE, B. ROSNER, AND J. YAN (2020): "Wilcoxon Rank-Based Tests for Clustered Data with R Package clusrank," *Journal of Statistical Software*, 96. Cited on pages 17, 18, 19, 20, 25, 26, 52, and 53.

JOLSON, M. A. AND G. L. ROSSOW (1971): "The Delphi Process in Marketing Decision Making," *Journal of Marketing Research*, 8, 443–448. Cited on page 9.

KAPLAN, T. R., L. CHOO, AND R. ZULTAN (2022): "Manipulation and (Mis)trust in Prediction Markets." *Management Science*. Cited on pages 3, 7, and 8.

KHALMETSKI, K., B. ROCKENBACH, AND P. WERNER (2017): "Evasive lying in strategic communication," *Journal of Public Economics*, 156, 59–72. Cited on page 10.

LOVALLO, D., T. KOLLER, R. UHLANER, AND D. KAHNEMAN (2020): "Your company is too risk-averse," *Harvard Business Review*, 98. Cited on page 2.

MACIEJOVSKY, B. AND D. V. BUDESCU (2013): "Markets as a structural solution to knowledge-sharing dilemmas," *Organizational Behavior and Human Decision Processes*, 120, 154–167. Cited on pages 7, 8, and 13.

——— (2020): "Too Much Trust in Group Decisions: Uncovering Hidden Profiles by Groups and Markets," *Organization Science*, 31, 1497–1514. Cited on pages 3, 7, 8, and 13.

MARETT, K. AND J. F. GEORGE (2012): "Barriers to Deceiving Other Group Members in Virtual Settings," *Group Decision and Negotiation*, 22, 89–115. Cited on page 10.

MATTOZZI, A. AND M. Y. NAKAGUMA (2022): "Public Versus Secret Voting in Committees," *Journal of the European Economic Association*, 21, 907–940. Cited on pages 2 and 10.

MAZAR, N., O. AMIR, AND D. ARIELY (2008): "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," *Journal of Marketing Research*, 45, 633–644. Cited on page 10.

MCDOWELL, M. AND P. JACOBS (2017): "Meta-analysis of the effect of natural frequencies on Bayesian reasoning," *Psychol Bull*, 143, 1273–1312. Cited on page 48.

MELLERS, B., L. UNGAR, J. BARON, J. RAMOS, B. GURCAY, K. FINCHER, S. E. SCOTT, D. MOORE, P. ATANASOV, S. A. SWIFT, T. MURRAY, E. STONE, AND P. E. TETLOCK (2014): "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," *Psychological Science*, 25, 1106–1115. Cited on pages 9 and 37.

NELSON, B. W. (1978): "Statistical manipulation of delphi statements: Its success and effects on convergence and stability," *Technological Forecasting and Social Change*, 12, 41–60. Cited on page 8.

PARENTÉ, F. AND J. ANDERSON-PARENTÉ (1987): *Delphi inquiry systems*, Chichester: Wiley, 129–156. Cited on page 9.

PEARSALL, M. J. AND V. VENKATARAMANI (2015): "Overcoming asymmetric goals in teams: the interactive roles of team learning orientation and team identification," *J Appl Psychol*, 100, 735–48. Cited on page 2.

PEETERS, R. AND L. WOLK (2017): "Eliciting interval beliefs: An experimental study," *PLoS One*, 12, e0175163. Cited on page 14.

——— (2018): "Elicitation of expectations using Colonel Blotto," *Experimental Economics*, 22, 268–288. Cited on page 14.

ROSNER, B., R. J. GLYNN, AND M. L. LEE (2003): "Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach," *Biometrics*, 59, 1089–98. Cited on pages 17, 18, 19, 20, 52, and 53.

ROTHSCHILD, D. AND R. SETHI (2016): "Trading Strategies and Market Microstructure: Evidence from a Prediction Market," *The Journal of Prediction Markets*, 10. Cited on page 2.

ROWE, G. AND G. WRIGHT (1996): "The impact of task characteristics on the performance of structured group forecasting techniques," *International Journal of Forecasting*, 12, 73–89. Cited on page 9.

——— (1999): "The Delphi technique as a forecasting tool: issues and analysis," *International Journal of Forecasting*, 15, 353–375. Cited on pages 3, 6, 9, and 13.

——— (2001): *Expert Opinions in Forecasting: The Role of the Delphi Technique*, Boston: Springer. Cited on pages 9 and 16.

ROWE, G., G. WRIGHT, AND F. BOLGER (1991): "Delphi: A reevaluation of research and theory," *Technological Forecasting and Social Change*, 39, 235–251. Cited on pages 6, 9, and 14.

ROWE, G., G. WRIGHT, AND A. MCCOLL (2005): "Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence," *Technological Forecasting and Social Change*, 72, 377–399. Cited on page 9.

SATOPÄÄ, V., M. SALIKHOV, P. TETLOCK, AND B. MELLERS (2021): "Bias, Information, Noise: The BIN Model of Forecasting," *Management Science*. Cited on pages 5, 28, 30, 34, and 55.

SATOPÄÄ, V., M. SALIKHOV, AND E. MORENO (2022): "BINtools," . Cited on pages 31 and 55.

SCHLAG, K. H. AND J. J. VAN DER WEELE (2015): "A method to elicit beliefs as most likely intervals." *Judgment & Decision Making*, 10. Cited on pages 4 and 51.

SERRA-GARCIA, M. AND U. GNEEZY (2021): "Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies," *American Economic Review*, 111, 3160–3183. Cited on page 10.

SERRA-GARCIA, M., E. VAN DAMME, AND J. POTTERS (2011): "Hiding an inconvenient truth: Lies and vagueness," *Games and Economic Behavior*, 73, 244–261. Cited on page 10.

STASSER, G. AND W. TITUS (1985): "Pooling of Unshared Information in Group Decision Making: Biased Information Sampling During Discussion," *Journal of Personality and Social Psychology*, 48, 1467–1478. Cited on page 48.

SUNSTEIN, C. R. (2005): "Group Judgments: Statistical Means, Deliberation, and Information Markets," *New York University Law Review*, 80, 962. Cited on page 3.

SUTTER, M. (2009): "Deception Through Telling the Truth?! Experimental Evidence from Individuals and Teams," *The Economic Journal*, 119, 47–60. Cited on page 10.

TESCHNER, F., D. ROTHSCHILD, AND H. GIMPEL (2017): "Manipulation in Conditional Decision Markets," *Group Decision and Negotiation*, 26, 953–971. Cited on page 8.

TETLOCK, P. E. (2005): *Expert Political Judgment: How Good Is It? How Can We Know?*, Princeton University Press. Cited on page 20.

TOMA, C. AND F. BUTERA (2009): "Hidden profiles and concealed information: strategic information sharing and use in group decision making," *Pers Soc Psychol Bull*, 35, 793–806. Cited on page 2.

VISSER, B. AND O. H. SWANK (2007): "On Committees of Experts," *The Quarterly Journal of Economics*, 122, 337–372. Cited on page 2.

VON MEIJENFELDT, F. A. B., P. HOGEWEG, AND B. E. DUTILH (2023): "A social niche breadth score reveals niche range strategies of generalists and specialists," *Nature Ecology & Evolution*, 7, 768–781. Cited on page 21.

WINTLE, B., S. MASCARO, F. FIDLER, M. MCBRIDE, M. BURGMAN, L. FLANDER, G. SAW, C. TWARDY, A. LYON, AND B. MANNING (2012): "The Intelligence Game: Assessing Delphi Groups and Structured Question Formats." . Cited on page 9.

WITTROCK, L. (2023): "Useful forecasting: belief elicitation for decision making," *Working Paper*. Cited on page 8.

WOLFERS, J. AND E. ZITZEWITZ (2004): "Prediction Markets," *Journal of Economic Perspectives*, 18, No. 2. Cited on page 2.

WOUDENBERG, F. (1991): "An evaluation of Delphi," *Technological Forecasting and Social Change*, 40, 131–150. Cited on pages 6, 9, and 14.

A Experimental details

A.1 Estimation experiment

A.1.1 Judgment task

The judgment task is framed in a gamified story to ease understanding and increase the engagement of experimental participants. In particular, the movement path is framed to be generated by a computer program. Group members are told that the computer program generates the movement path of a ladybird on a map. They need to estimate the chance that the ladybird reaches the target area at the end of its path after ten steps. The chance is to be expressed in frequencies, such as 75 in 100, to facilitate better statistical reasoning (McDowell and Jacobs, 2017). As information, each sees a movement path generated by a distinct past execution of the program, compare Figure 10. For each round, group members are informed that a distinct computer program generates the movement paths, and consequently, the underlying probability of reaching the target might be different.

This judgment task amalgamates key features of the hidden-profile paradigm (Stasser and Titus, 1985), a benchmark task of group interaction in psychology, with additional aspects enabling more detailed analysis and broader real-world references. Common features comprise experimenter control over the information needed to solve the problem and its distribution across group members, i.e., participants resemble heterogeneous and complementary experts with the degree of expertise induced by the experiment. Consequently, the more private information is shared during group interaction, the more accuracy can be achieved in the group estimate. However, going beyond the standard hidden profile paradigm, this task incorporates uncertainty and requires probabilistic judgments. This comes with four major advantages. First, estimates can be evaluated based on a statistical measure of accuracy. Second, the statistical measure of accuracy enables incentives that continuously reward accuracy and hidden agenda achievement. Third, participants' behavior can be evaluated against theoretically optimal behavior given the information they have at hand. Finally, probabilistic judgment caters to real-world applications that go far beyond the motivating examples presented in the introduction.

Figure 10: Example of information received by group members (Person A, B, C, and D)
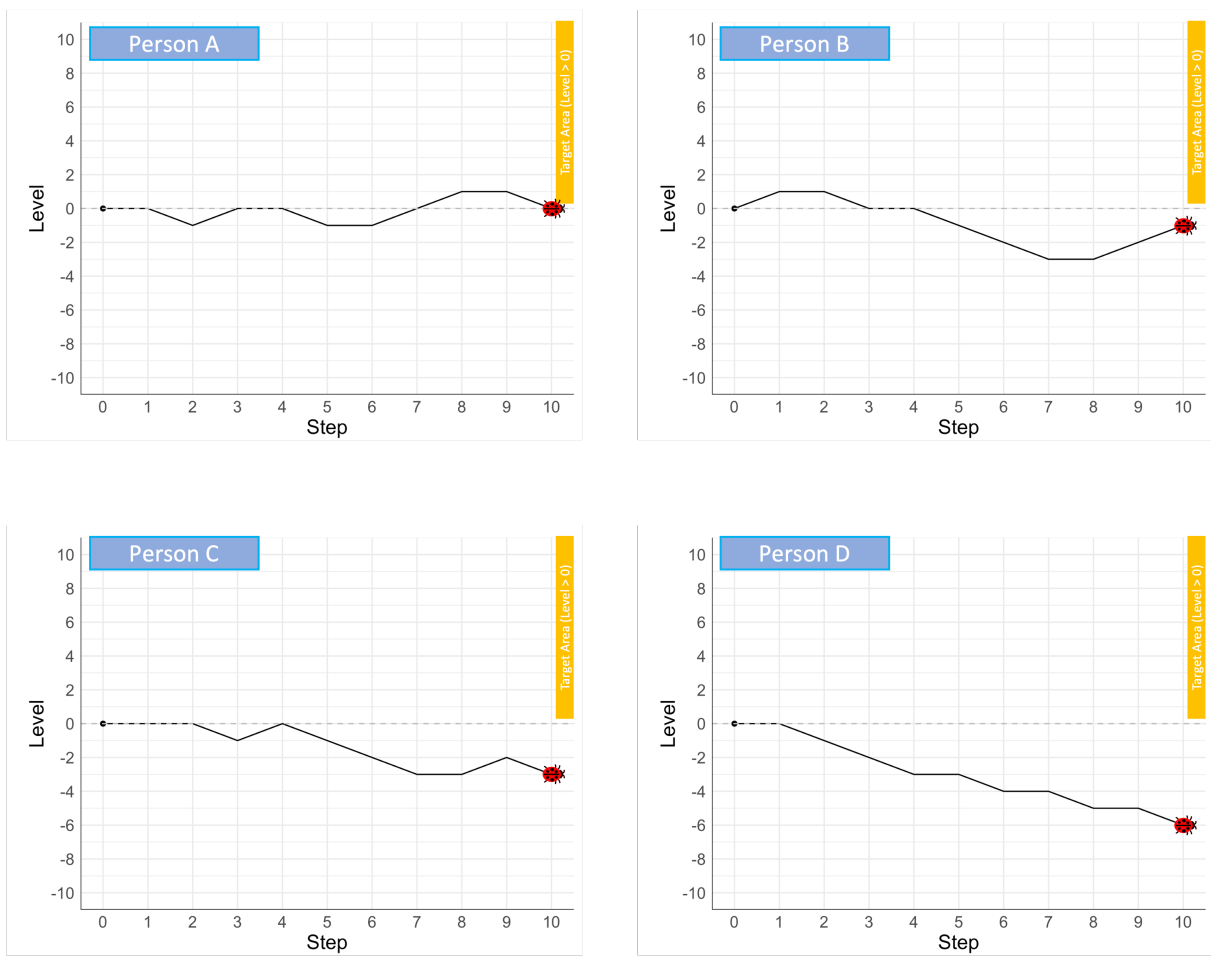
A.1.2 Incentives

All participants receive a show-up fee of €5 and performance-based remuneration for solving the task as accurately as possible. Additionally, to create a situation of hidden agendas, two group members can earn a supplementary individual bonus in the hidden agenda treatment arm by influencing the group's judgment in a particular direction.

For each of the ten rounds of the judgment task, all groups receive an **accuracy-based bonus**, calculated based on the binarized quadratic scoring rule (Hossain and Okui, 2013). The group bonus is split equally among all four group members. This ensures that individuals are incentivized to pursue the most accurate group estimate irrespective of their risk preferences. Specifically, the group bonus in a given round amounts to €6 with a certain chance and €0 otherwise. The more accurate the group estimate, the greater the chance the group bonus is €6. To illustrate the calculation of the group bonus, let $Y_r \epsilon(0,1)$ be the binary event that the movement path does (not) reach the target in round r. $Y_r = 1$ if the movement path reaches the target in round r, i.e. $\sum_{t=1}^{10} X_t > 0$, and 0 otherwise. Moreover, p_r is the group estimate of the chance that the movement path reaches the target area in a particular round r. Following the binarized quadratic scoring rule, then the group bonus in round r is €6 if $Y_r = 1$ and a random number $Y \sim U(0,1) > (1 - p_r)^2$ or if $Y_r = 0$ and a random number $Y \sim U(0,1) > p_r^2$.[23] By way of illustration, consider the true chance that the movement path reaches the target is $p^* = 0.7$, and the group estimated the chance at $p_r = 0.6$. In this case, 70% of actual realizations will reach the target, and 30% will not. Consequently, the chance of winning the group bonus is $0.7*(1-(1-0.6)^2)+0.3*(1-0.6^2)$, which is $0.7*0.84+0.3*0.64 = 0.78$ In other words, the chance of winning the group bonus is 78% in total, i.e. in expectation €4.68 group bonus or in other words €1.17 bonus per person.

In hidden agenda situations, in addition to the accuracy-based bonus, two out of four group members may each receive an **individual hidden agenda bonus** of €1.50 per round based on a binarized scoring. Their hidden agenda is to drive the group estimate to 1 or 0, i.e., up or down. Whether it is one or the other direction is determined randomly in each round. The direction of the hidden agenda is always the same for the two group members with a hidden agenda in a given round. The hidden agenda bonus is designed

---

[23]The payoff relevant realizations of the movement path Y_r have been generated for all rounds $r = 1, ..., 10$ once before the experiment and are kept constant for all groups in the experiment.

such that it is always best for expected payoff-maximizing participants to follow their hidden agenda as much as possible. If following their hidden agenda, the decrease in the chance of earning the group bonus is overcompensated by gains in the chance of earning the individual hidden agenda bonus. Precisely, the chance of winning the hidden agenda bonus in round r is $(1 - p_r)^2$ if the hidden agenda is to drive the estimate to 0, i.e., down, and p_r^2 if the hidden agenda is to drive the estimate to 1, i.e., up.

A.2 Decision experiment

A.2.1 Incentives

All participants in the decision experiment receive a show-up fee of €10 and performance-based remuneration depending on their stated confidence intervals. The latter follows the most likely interval method (Schlag and van der Weele, 2015), i.e., a stated confidence interval with lower bound L and upper bound U translates into a payment S:

$$S(L, U, p^*) = \begin{cases} 10(1 - \frac{U-L}{100}) & \text{if } 100p^* \in [L, U] \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

depending on the respective true probability p^*. Note that estimates generated in the estimation experiment are naturally bounded by 0 and 100 in frequency terms. Further, the width of the stated confidence interval ranging from lower bound L to upper bound is $U - L$, which is at most 100. Consequently, the wider the stated confidence interval, the smaller the bonus paid in case the confidence interval contains $100p^*$, ultimately approaching 0 if the confidence interval spans the entire interval of possible values. I implement performance-based remuneration as a random lottery incentive, i.e., after participants stated their confidence intervals on all 40 estimates, one confidence interval is chosen at random for payout according to the above-mentioned most likely interval payment rule. This method is designed to encourage participants to treat each interval with equal, high care.

B Robustness checks

B.1 Accuracy

While the main analysis of accuracy focuses on absolute errors, accuracy can also be quantified using alternative metrics. Accordingly, I report test results based on squared errors (Figure 11) and Brier Scores (Figure 12). Furthermore, I report test results varying in the strictness of considering the dependence of judgments made within the same group. Two of the main results reported in Section 4 are robust in all alternative tests: First, hidden agendas decrease accuracy only for Delphi groups, and second with hidden agendas FTF groups are more accurate. Without hidden agendas, however, Delphi groups are no longer statistically more accurate than FTF groups in the robustness test, which is most strict concerning the potential dependence of observations.

Figure 11: Squared errors of group judgments across experimental conditions
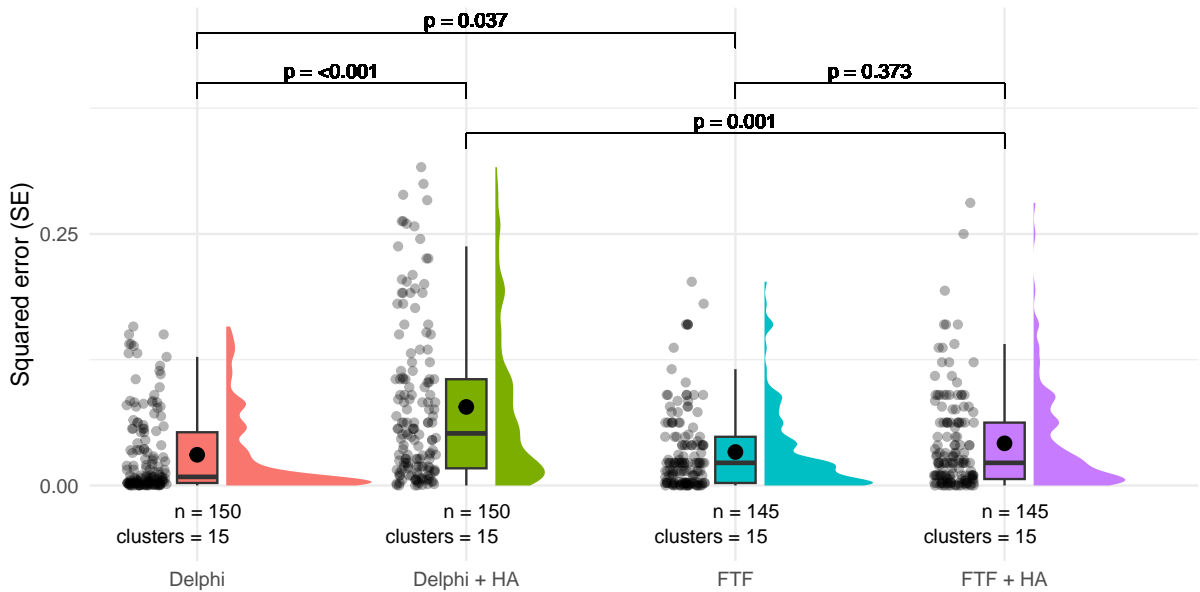
Notes: P-values from two-sided Wilcoxon rank sum tests (Rosner et al., 2003; Jiang et al., 2020). Bar in boxplot = median; dot in boxplot = mean.

Squared errors (SE), just like absolute errors juxtapose group judgments ($p_{g,r}$) to true probabilities (p_r^*), but consider the squared difference:

$$SE_{g,r} = (p_{g,r} - p_r^*)^2 \tag{11}$$

Figure 12: Brier Scores of group judgments across experimental conditions
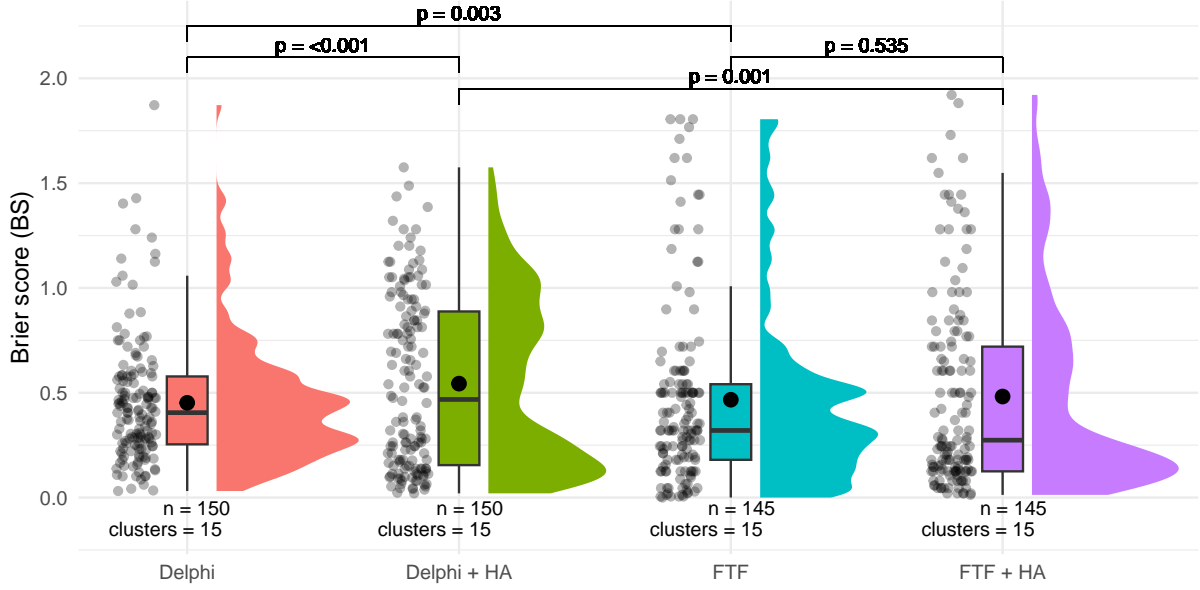
Notes: P-values from two-sided Wilcoxon rank sum tests (Rosner et al., 2003; Jiang et al., 2020). Bar in boxplot = median; dot in boxplot = mean.

Consequently, the distribution of squared errors is narrower compared to absolute errors. However, the rank of observations remains unchanged. Therefore, test results are identical to those reported in Section 4.

Brier Scores (BS) are calculated based on group judgments ($p_{g,r}$) and actual outcomes (Y_r), i.e., binary outcomes (0 or 1 for hit or missed target, respectively) drawn randomly according to the true underlying probabilities (p_r^*):

$$BS_{g,r} = (Y_r - p_{g,r})^2 + ((1 - Y_r) - (1 - p_{g,r}))^2 \tag{12}$$

Testing for differences in the distributions of Brier Scores across treatment conditions, I find the same results as with AEs and SEs. This is, however a stronger result, as Brier Scores are noisier than absolute and squared errors, which holds especially with a small number of distinct judgment tasks, like the ten different rounds in the estimation experiment. For illustration, consider that the underlying true probabilities are $[0.1, 0.2, 0.3, 0.4, 0.45, 0.55, 0.6, 0.7, 0.8, 0.9]$ and the accordingly drawn actual outcomes, which are used to calculate Brier scores, are $[0, 1, 0, 0, 1, 0, 1, 1, 1, 1]$.

Varying the strictness of considering the dependence of judgments made within the same group, I report results of Wilcoxon rank sum tests for average AE per group (most

Table 3: Wilcoxon rank sum tests on accuracy with different considerations of dependence of observations

| | Interaction | Hidden Agendas | | Interaction |
	FTF vs. Delphi	FTF vs. FTF^{HA}	Delphi vs. Delphi^{HA}	FTF^{HA} vs. Delphi^{HA}
Wilcoxon rank sum test on AE, average per group				
n	30	30	30	30
p-value	0.115	0.52	<0.001	0.001
Clustered Wilcoxon rank sum test on AEs (Rosner, Glynn & Lee 2003)				
n	295	290	300	295
cluster	30	30	30	30
p-value	0.037	0.373	<0.001	0.001

strict), in Table 3. The most strict scenario posits that groups make 10 judgments over the 10 rounds of the task, but in this way generate only one independent observation. Reporting test results only using average AE per group takes this into account. The clustered Wilcoxon rank sum tests reported in Section 4 are less strict. They actively control for the degree of dependence of observations within the same group. The most strict approach confirms the main results obtained through clustered Wilcoxon rank sum tests, except for Delphi groups being more accurate without hidden agendas. FTF and Delphi groups' accuracy becomes statistically indistinguishable in this case.

B.2   Trustworthiness

Similar to accuracy, I vary the strictness of considering the dependence of observations generated in the decision experiment. The main analysis reports results of clustered Wilcoxon signed rank tests that actively control for the dependence of confidence intervals stated by the same decision-maker. Obtained test results are robust towards considering the most strict scenario, positing that a decision-makers evaluating ten judgments from a given condition of the estimation experiment only generates one independent observation. Accounting for this, Table 4 reports results based on the average confidence interval width per decision-maker and condition in the estimation experiment.

Table 4: Wilcoxon signed rank tests on trust with different considerations of dependence of observations

	Interaction	Hidden Agendas	
	FTF vs. Delphi	FTF vs. FTF^{HA}	Delphi vs. Delphi^{HA}
Wilcoxon signed rank test on interval widths, average per participant			
n	100	100	100
test statistic	866	286.5	280
p-value	0.028	0.001	0.001
Clustered Wilcoxon signed rank test on interval widths (Datta & Satten 2008)			
n	1000	1000	1000
cluster	50	50	50
p-value	<0.001	<0.001	<0.001

C Modelling framework

C.1 BIN-model estimation

The implementation of the BIN-model largely follows the econometric procedures of Satopää et al. (2021). For details on these procedures, such as the underlying likelihood function, I refer the reader to the original paper and the documentation of the statistical companion package BINtools (Satopää et al., 2022). In the following, I focus on outlining adaptions to the original procedures I made to accommodate better the data from the estimation experiment of this study.

First, I exploit the fact that within my experimental design, the true probabilities for the ten distinct forecasts to be made by each group are known. In particular, each group faces judgment tasks with objectively true probabilities $\mathbb{P}(\sum_{t=1}^{10} X_t > 0) = \{0.1, 0.2, 0.3, 0.4, 0.45, 0.55, 0.6, 0.7, 0.8, 0.9\}$. With this knowledge, I can fix $\mu^* = 0$, i.e., the mean of the normal distribution from which actual true probabilites in the model are drawn. [24] Second, I increase the numbers of iterations `warmup` and `iter`, and I set estimation control parameters for `adapt_delta`, `stepsize`, and `max_treedepth` according to

---

[24] The parameter is fixed in the Stan source code of the R BINtools package by restricting $-0.01 < \mu^* < 0.01$.

Stan estimation feedback provided by BINtools. Third, I perform an iterative grid-search on the initial values of model parameters μ_0, μ_1, γ_0, γ_1, δ_0, ρ_0, δ_1, ρ_1, and ρ_{01}, to reduce the number of divergent transitions after warmup.

D   Analysis of communication protocols

To analyze emergent communication patterns across conditions of the estimation experiment group and round specific transcripts were created. For Delphi groups, I extracted these from the written reasoning provided by group members alongside their first individual estimate. For FTF groups, research assistants who were unaware of the research questions underlying this paper, transcribed the video recordings of respective experimental sessions.[25] Subsequently, the transcripts were coded according to the coding scheme on characteristics of information sharing shown in Online Appendix E.[26]

---

[25]The research assistants used the automatically generated transcripts of the zoom video call as a starting point and corrected parts of the communication that had been incorrectly transcribed.

[26]Coding and analysis of communication protocols according to the preregistered coding schemes on the general impression of group interaction and on quantifiable characteristics of group interaction have been postponed to future research.

Table 5: OLS estimation results for linear probability models on information revelation across conditions of the estimation experiment

	target	target_lie	target_truth	level	level_lie	level_truth
(Intercept)	0.047***	0.003	0.950***	0.112***	0.033**	0.855***
	(0.013)	(0.009)	(0.015)	(0.015)	(0.012)	(0.017)
delphi_ha	0.227***	0.127***	−0.353***	0.382***	0.113***	−0.495***
	(0.018)	(0.013)	(0.021)	(0.021)	(0.016)	(0.024)
ftf	0.103***	0.003	−0.107***	0.078***	−0.015	−0.063**
	(0.018)	(0.013)	(0.021)	(0.021)	(0.016)	(0.024)
ftf_ha	−0.027	0.082***	−0.055**	−0.078***	0.148***	−0.070**
	(0.018)	(0.013)	(0.021)	(0.021)	(0.016)	(0.024)
Num.Obs.	2400	2400	2400	2400	2400	2400

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	some_steps	some_steps_lie	some_steps_truth	all_steps	all_steps_lie	all_steps_truth
(Intercept)	0.035*	0.073***	0.892***	0.285***	0.035***	0.680***
	(0.015)	(0.013)	(0.018)	(0.019)	(0.010)	(0.019)
delphi_ha	0.162***	0.128***	−0.290***	0.422***	0.045**	−0.467***
	(0.021)	(0.019)	(0.026)	(0.027)	(0.015)	(0.027)
ftf	0.295***	−0.028	−0.267***	0.335***	−0.015	−0.320***
	(0.021)	(0.019)	(0.026)	(0.027)	(0.015)	(0.027)
ftf_ha	0.117***	0.097***	−0.213***	−0.020	0.122***	−0.102***
	(0.021)	(0.019)	(0.026)	(0.027)	(0.015)	(0.027)
Num.Obs.	2400	2400	2400	2400	2400	2400

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Each column represents a distinct dependent variable, a binary indicator of information revelation. *(1) target:* not revealing whether the target has been reached, *(2) target_lie:* misstating whether the target has been reached, *(3) target_truth:* correctly stating whether the target has been reached, *(4) level:* not revealing the final level of the movement path, *(5) level_lie:* misstating the final level of the movement path, *(6) level_truth:* correctly stating the final level of the movement path, *(7) some_steps:*, not revealing any particular steps of the movement path *(8) some_steps_lie:*, revealing one or more particular steps of the movement path, of which at least one is misstated *(9) some_steps_truth:* *revealing one or more particular steps of the movement path, of which none is misstated, (10) all_steps:* not revealing all ten particular steps of the movement path, *(11) all_steps_lie: revealing all steps of the movement path, of which at least one is misstated, (12) all_steps_truth:* revealing all steps of the movement path, of which none is misstated.

Online Appendix

E   Coding scheme for video recordings

E.1   Preregistered coding scheme

Questions on the general impression of group interaction

- Did the group members follow a structure to discuss and generate a group estimate? (5-point Likert scale, strongly agree to strongly disagree)

  - Guidance: Did they take steps that were the same in multiple rounds? Steps could e.g., be an open discussion where everybody presents their arguments, a step where everybody presents their individual estimate, an aggregation step where individual estimates are averaged to generate a group estimate, etc.

  - Please briefly describe the structure followed by the group and the steps taken, if applicable. (open text)

- Please rate the overall influence of person A, B, C, and D throughout the group discussions. (5-point Likert scale, very influential to very uninfluential)

- Is there anything else you found striking about the group interaction? (open text)

Quantifiable characteristics of group interaction (coded for each of the 10 estimation rounds separately)

- Words spoken per person A, B, C, and D (extracted from automatically generated video conference transcripts)

- Number of numerical estimates uttered per person A, B, C, and D (extracted from automatically generated video conference transcripts)

  - Guidance: Numerical estimates could be in the form of frequencies, e.g. 10 in 100, the form of probabilities, e.g. 0.1, or the form of percentages, e.g. 10%.

- Number of qualitative/verbiage estimates uttered per person A, B, C, and D (extracted from automatically generated video conference transcripts)

– Guidance: Qualitative/verbiage estimates are e.g. very likely, very probable, impossible, almost certain, etc.

- Number of mentions of "hidden agendas" or "manipulation" per person A, B, C, and D (extracted from automatically generated video conference transcripts)

  – Guidance: Qualitative/verbiage estimates are e.g. very likely, very probable, impossible, almost certain, etc.

- Number of arguments uttered per person A, B, C, and D

  – Guidance: An argument usually entails a premise: e.g. My info indicates ..., and a conclusion: e.g. therefore I believe the chance to be ...

  – Repetitions are counted the same way as genuine arguments

- Number of confirmations uttered per person A, B, C, and D

  – Guidance: A confirmation could be: "I agree with the argument of ...", "I think the point of ... is true", etc.

- Number of confirmations from other persons received per person A, B, C, and D

  – Guidance: A confirmation could be: "I agree with the argument of ...", "I think the point of ... is true", etc.

- Number of qualifications uttered per person A, B, C, and D

  – Guidance: A qualification could be: "I do partly agree with ..., but I think ...", etc.

- Number of qualifications from other persons received per person A, B, C, and D

  – Guidance: A qualification could be: "I do partly agree with ..., but I think ...", etc.

- Number of direct rebuttals/attacks on other persons per person A, B, C, and D

– Guidance: A direct rebuttal/attack could be: "I do disagree with the argument of ...", "I think the point of ... cannot be true", etc.

- Number of direct rebuttals/attacks received per person A, B, C, and D

  – Guidance: A direct rebuttal/attack could be: "I do disagree with the argument of ...", "I think the point of ... cannot be true", etc.

- Please rate the state of agreement on the final group estimate. (5-point Likert scale, strong agreement to strong disagreement)

E.2 Coding scheme extension post data-collection

Characteristics of information sharing (coded for each of the 10 estimation rounds separately in Delphi and FTF treatments)

- Did person A, B, C and D state whether their movement path reached the target? (no, yes truthfully, yes but not truthfully)

- Did person A, B, C, and D reveal the final level of their movement path? (no, yes truthfully, yes but not truthfully)

- Did person A, B, C, and D reveal information on one or more values (-1,0,1) of individual steps of their movement path? (no, yes truthfully, yes but not truthfully)

  – Guidance: The answer is yes e.g. if person A states "The ladybird goes down in the first step ..." or "For me the last step is 0..."

- Did person A, B, C, and D reveal information on all 10 values (-1,0,1) of individual steps of their movement path truthfully? (no, yes truthfully, yes but not truthfully)

  – Guidance: The answer is yes e.g. if person B states "I have 4 ups, 3 downs and 3 straights..." or person B states "My ladybird goes -1,1,0,0,1,1,-1,0,-1,1 ..." and this information matches the movement path seen by person in B in the given round

In cases of doubt, the following rules were applied. If there was doubt about whether a piece of information has or has not been revealed, it was considered as revealed. If there was doubt whether the revealed information was truthful or not, it was considered truthful. If the revealed information allowed to infer other information, the other information was considered revealed accordingly, e.g., a stated final level of 2 implies stating that the target has been reached.

F Experimental instructions

F.1 Estimation experiment

In the following, I provide screenshots of the estimation experiment and instructions provided to the participants. The screenshots follow the chronological order of the experiment. Where applicable, the screenshots and accompanying notes highlight differences across treatments.

F.1.1 Welcome information

Figure 13: Opening screen of the estimation experiment
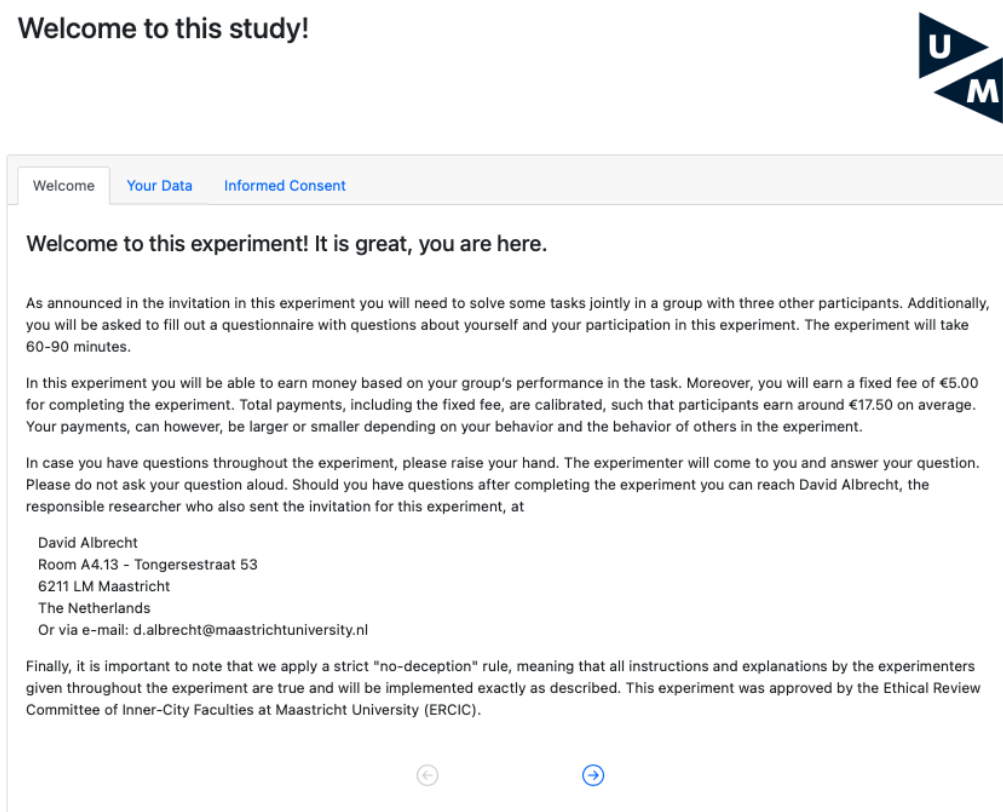
Figure 14: Information on data handling in the estimation experiment

Notes: This screen is part of a face-to-face treatment, and participants are informed that the experiment will be video recorded. There is no video recording in the Delphi treatments, and the respective information is not displayed to participants.

Figure 15: Informed consent in the estimation experiment

Welcome to this study!

| Welcome | Your Data | Informed Consent |

Your consent to participate

I hereby give permission for my personal data to be used for this research project. I have had enough time to decide whether I want to participate in the experiment. I know that participation is voluntary, and I know that I can decide to withdraw from the experiment at any time. I do not have to justify such a decision to withdraw. If I withdraw, I will forfeit any payments.

I understand that personal data collected throughout the experiment will be stored on secure servers of Maastricht University and only be available to the research team as outlined before. Only fully anonymized data may be made available outside the research team. Moreover, I give permission for the researcher to use my anonymised responses in subsequent experiments.

Participate in this experiment

F.1.2 Introduction to the experiment

Figure 16: Introduction to the group judgment task
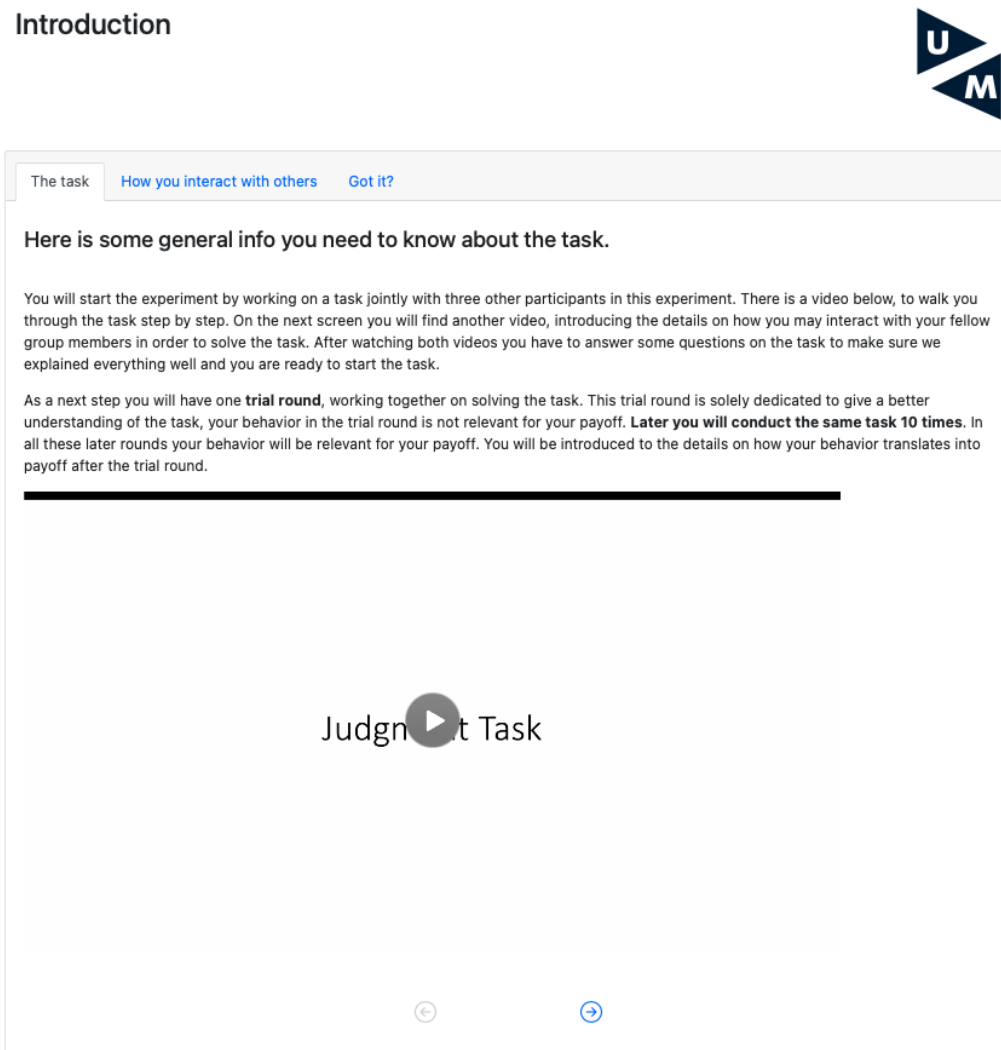
Notes: The introduction video on the judgment task can be found in the replication package.

Figure 17: Introduction to the group interaction format

Notes: Depending on the treatment the video introduced FTF or Delphi interaction. Both respective introduction videos on the interaction format can be found in the replication package.

Figure 18: Check of understanding in FTF treatments

Notes: Upon hitting the "Check answers" button, participants are informed which questions still contain incorrect answers. They can revisit the introduction videos and try to re-answer the questions. Only, after all questions are answered correctly participants may continue in the experiment. The number of attempts is recorded.

Figure 19: Check of understanding in Delphi treatments

Notes: Upon hitting the "Check answers" button, participants are informed which questions still contain incorrect answers. They can revisit the introduction videos and try to re-answer the questions. Only, after all questions are answered correctly participants may continue in the experiment. The number of attempts is recorded.

F.1.3   Judgment task

Figure 20: Introduction / repetition of payment information in FTF treatment without hidden agendas

Notes: The corresponding introduction video on the bonus payments can be found in the replication package.

Figure 21: Introduction / repetition of payment information to participants without hidden agenda in FTF treatment with hidden agenda agendas

Round 1 out of 10

Time left to complete this page: **9:36**

| Your payoff | Your Info | Solving the task |

This is how your behavior in the task translates into your payoff.

Your only objective is to reach a group estimate that is as accurate as possible.

In particular, your group may earn a bonus of €6.00 for this round. The bonus will be split equally among group members, in other words you may earn €1.50 for yourself.

The chance of earning the group bonus depends on the accuracy of your group's estimate to the judgment task. The more accurate the larger the chance. In this way, **it will always be best for you to strive for a group estimate, which you believe is the most accurate given the available information**.

Be wary: **two of your fellow group members have a hidden agenda.** They do not only receive their share of the group bonus but also earn money for driving the estimate as close to *0 in 100* or *100 in 100* as possible. The hidden agenda for both of them goes in the same direction in a given round. Whether it is *0 in 100* or *100 in 100* is decided randomly in each round.

Bonuses ▶

If you want to know all details about how the group bonus and the hidden agenda bonus are calculated you can click here.

Notes: The corresponding introduction video on the bonus payments can be found in the replication package.

Figure 22: Introduction / repetition of payment information to participants with hidden agenda in FTF treatment with hidden agenda agendas

Round 1 out of 10

Time left to complete this page: **9:14**

| Your payoff | Your Info | Solving the task |

This is how your behavior in the task translates into your payoff.

You have a hidden agenda. Consequently, you have two objectives. Your first aim is to work on your hidden agenda as outlined below.

Your hidden agenda in this round is to drive the estimate as close to *100 in 100* as possible. You may earn a hidden agenda bonus of €1.50 for yourself. The chance of winning this hidden agenda bonus increases the closer you drive the estimate in the prescribed direction.

Second, you are still part of a group, which collectively follows the aim to reach a group estimate that is as accurate as possible. In particular, your group may earn a bonus of €6.00 for this round. The bonus will be split equally among group members, in other words you may earn €1.50 for yourself.

The chance of earning the group bonus depends on the accuracy of your group's estimate to the judgment task. The more accurate the larger the chance. In this way, from the group's perspective it will always be best to generate an estimate, which the group believes is the most accurate given the available information. **For yourself, it will always be best to follow your hidden agenda as much as possible.** Your bonus based on the hidden agenda outweighs your share of the group accuracy bonus.

Bonuses

Notes: The corresponding introduction video on the bonus payments can be found in the replication package.

Figure 23: Introduction / repetition of payment information to participants without hidden agenda in Delphi treatment with hidden agenda agendas

Round 1 out of 10

Your payoff | Your Info | Solving the task

Your objectives and how they translate into your payoff

Your only objective is to reach a group estimate that is as accurate as possible.

In particular, your group may earn a bonus of €6.00 for this round. The bonus will be split equally among group members, in other words you may earn €1.50 for yourself.

The chance of earning the group bonus depends on the accuracy of your group's estimate to the judgment task. The more accurate the larger the chance. In this way, **it will always be best for you to strive for a group estimate, which you believe is the most accurate given the available information**.

Be wary: **two of your fellow group members have a hidden agenda.** They do not only receive their share of the group bonus but also earn money for driving the estimate as close to *0 in 100* or *100 in 100* as possible. The hidden agenda for both of them goes in the same direction in a given round. Whether it is *0 in 100* or *100 in 100* is decided randomly in each round.

Bonuses

If you want to know all details about how the group bonus and the hidden agenda bonus are calculated you can click here.

Notes: The corresponding introduction video on the bonus payments can be found in the replication package.

Figure 24: Introduction / repetition of payment information to participants with hidden agenda in Delphi treatment with hidden agenda agendas

Your payoff   Your Info   Solving the task

Your objectives and how they translate into your payoff

You have a hidden agenda. Consequently, you have two objectives. Your first aim is to work on your hidden agenda as outlined below.

Your hidden agenda in this round is to drive the estimate as close to *0 in 100* as possible. You may earn a hidden agenda bonus of €1.50 for yourself. The chance of winning this hidden agenda bonus increases the closer you drive the estimate in the prescribed direction.

Second, you are still part of a group, which collectively follows the aim to reach a group estimate that is as accurate as possible. In particular, your group may earn a bonus of €6.00 for this round. The bonus will be split equally among group members, in other words you may earn €1.50 for yourself.

The chance of earning the group bonus depends on the accuracy of your group's estimate to the judgment task. The more accurate the larger the chance. In this way, from the group's perspective it will always be best to generate an estimate, which the group believes is the most accurate given the available information. **For yourself, it will always be best to follow your hidden agenda as much as possible.** Your bonus based on the hidden agenda outweighs your share of the group accuracy bonus.

Bonuses

If you want to know all details about how the group bonus and the hidden agenda bonus are calculated you can click here.

Notes: The corresponding introduction video on the bonus payments can be found in the replication package.

Figure 25: Individual information on the judgment task in FTF treatments

Time left to complete this page: **9:23**

Your payoff | Your Info | Solving the task

Your Info

Below you can see your own private information. You see the path Kara took when the program was executed once in the past.

Everybody in your group receives a path from a distinct past execution of the program. Consequently, the path you see is in all likelihood different from what others in your group see.
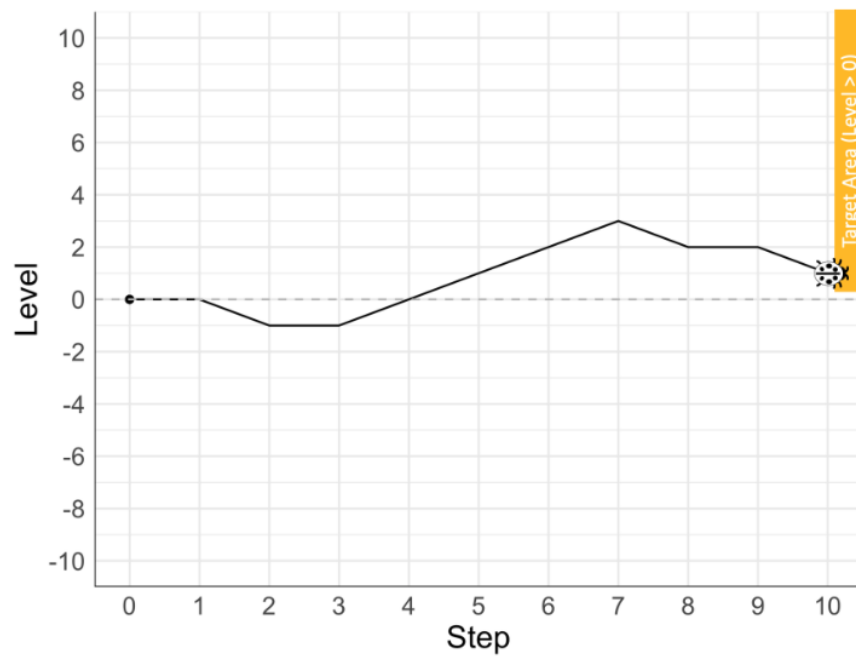
Figure 26: Individual information on the judgment task in Delphi treatments
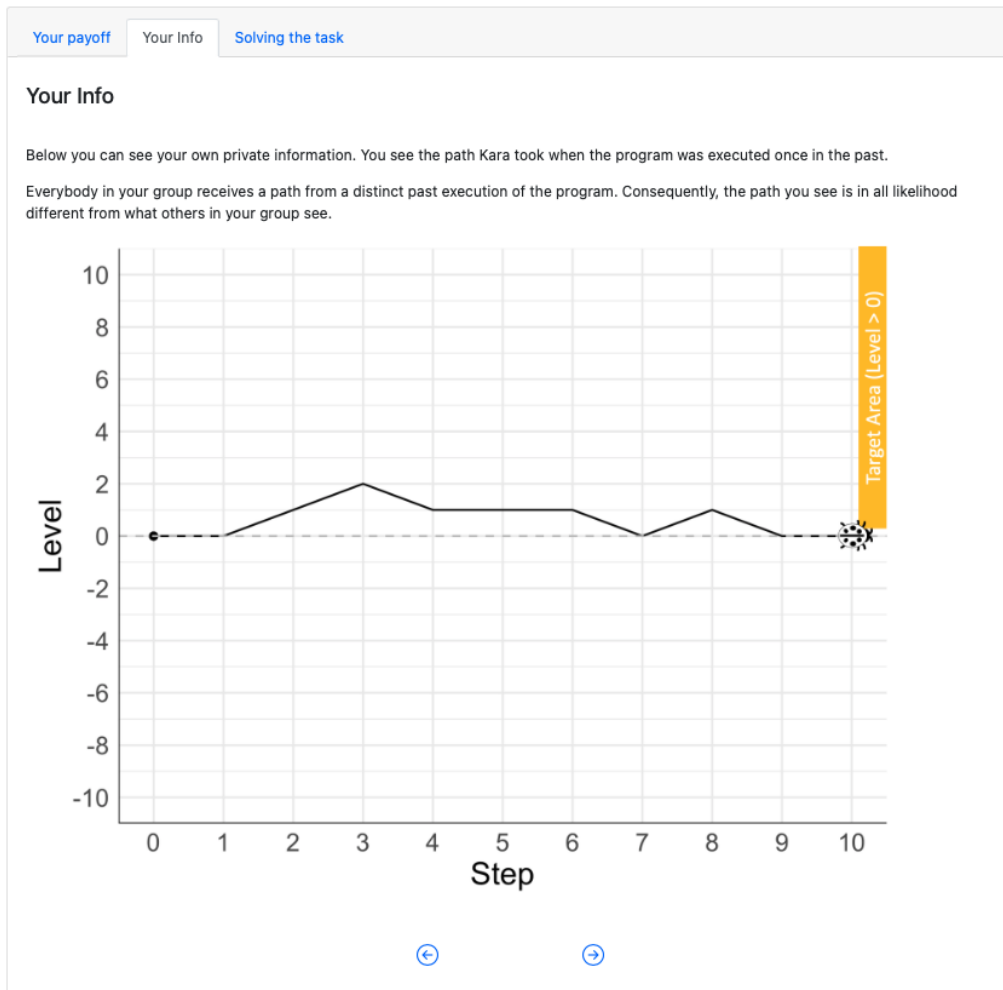
Figure 27: Input screen for group judgments in FTF treatments for participants without hidden agendas

Round 1 out of 10

Time left to complete this page: **9:00**

Your payoff   Your Info   Solving the task

Estimate

You can now **turn to your group in the video conference** on the right side of your screen in order to discuss your information and derive a group estimate.

Which estimate did you agree on during the group discussion?

[   ] in 100

Send estimate

Figure 28: Input screen for group judgments for participants with hidden agenda in FTF treatment with hidden agendas

Round 1 out of 10

Time left to complete this page: **8:43**

Your payoff   Your Info   Solving the task

Estimate

You can now **turn to your group in the video conference** on the right side of your screen in order to discuss your information and derive a group estimate.

Remember, your hidden agenda in this round is to drive the estimate as close to *100 in 100* as possible.

Which estimate did you agree on during the group discussion?

[   ] in 100

Send estimate

Figure 29: Input screen for first individual judgments for participants without hidden agenda in Delphi treatments

Round 1 out of 10

Your payoff   Your Info   Solving the task

First estimate

At first, all members of your group will give their individual estimate of the chance that Kara will reach the target area, when the program is executed for the next time. What is your estimate?

☐ in 100

Along the estimate you may also outline your reasoning behind the estimate in a brief text. Both your estimate and the reasoning will be presented anonymously to your fellow group members in the next step. What is your reasoning?

My estimate is based on the following line of thought ...

Send estimate and reasoning

Figure 30: Input screen for first individual judgments for participants with hidden agenda in Delphi treatments with hidden agendas

Round 1 out of 10

U
M

| Your payoff | Your Info | Solving the task |

First estimate

At first, all members of your group will give their individual estimate of the chance that Kara will reach the target area, when the program is executed for the next time. What is your estimate?

Remember, your hidden agenda in this round is to drive the estimate as close to *0 in 100* as possible.

[   ] in 100

Along the estimate you may also outline your reasoning behind the estimate in a brief text. Both your estimate and the reasoning will be presented anonymously to your fellow group members in the next step. What is your reasoning?

My estimate is based on the following line of thought ...

Send estimate and reasoning

←

Figure 31: Feedback screen after first individual judgments for participants without hidden agenda in Delphi treatments

Figure 32: Feedback screen after first individual judgments for participants with hidden agenda in Delphi treatments with hidden agendas

Figure 33: Feedback screen after second individual judgments for participants in Delphi treatments

Round 1 out of 10

Your payoff   Your Info   Solving the task

Done!

You completed round 1.

Below you can see your aggregated group estimate, and the second estimates of yourself and your fellow group members.

Please click the blue button to continue.

Aggregate group estimate

21.25 in 100

Your own estimate

0 in 100

Estimates of your fellow group members

Person A: 45 in 100
Person B: 35 in 100
Person C: 5 in 100

Continue to next round

F.1.4 Closing survey

Figure 34: Closing survey - personal characteristics

Figure 35: Closing survey - honesty module

Questionnaire

Now, please read the following statements and indicate how much you agree with each of them.

If I want something from a person I dislike, I will act very nicely toward that person in order to get it.

strongly disagree ○ ○ ○ ○ ○ strongly agree

If I knew that I could never get caught, I would be willing to steal a million euros.

strongly disagree ○ ○ ○ ○ ○ strongly agree

I wouldn't use flattery to get a raise or promotion at work, even if I thought it would succeed.

strongly disagree ○ ○ ○ ○ ○ strongly agree

I would be tempted to buy stolen property if I were financially tight.

strongly disagree ○ ○ ○ ○ ○ strongly agree

If I want something from someone, I will laugh at that person's worst jokes.

strongly disagree ○ ○ ○ ○ ○ strongly agree

I would never accept a bribe, even if it were very large.

strongly disagree ○ ○ ○ ○ ○ strongly agree

I wouldn't pretend to like someone just to get that person to do favors for me.

strongly disagree ○ ○ ○ ○ ○ strongly agree

I'd be tempted to use counterfeit money, if I were sure I could get away with it.

strongly disagree ○ ○ ○ ○ ○ strongly agree

Figure 36: Closing survey - experience during experiment

F.2   Decision experiment

In the following I provide screenshots of the decision experiment and instructions provided to the participants. The screenshots follow the chronological order of the experiment.

F.2.1 Welcome information

Figure 37: Opening screen of the decision experiment

Figure 38: Information on data handling in the estimation experiment

Welcome to this study!

Welcome | **Your Data** | **Informed Consent**

Welcome to this experiment! It is great, you are here.

As announced in the invitation in this experiment you will need to solve some tasks. Additionally, you will be asked to fill out a questionnaire with questions about yourself and your participation in this experiment. The experiment will take around 30-45 minutes.

In this experiment you will be able to earn money based on your performance in the tasks. Moreover, you will earn a fixed fee of €10.00 for completing the experiment. Total payments, including the fixed fee, are calibrated, such that participants earn around €15.00 on average. Your payments can however be larger or smaller depending on your behavior and the behavior of others.

In case you have questions throughout the experiment, please raise your hand. The experimenter will come to you and answer your question. Please do not ask your question aloud. Should you have questions after completing the experiment you can reach David Albrecht, the responsible researcher who also sent the invitation for this experiment, at

David Albrecht
Room A4.13 - Tongersestraat 53
6211 LM Maastricht
The Netherlands
Or via e-mail: d.albrecht@maastrichtuniversity.nl

Finally, it is important to note that we apply a strict "no-deception" rule, meaning that all instructions and explanations by the experimenters given throughout the experiment are true and will be implemented exactly as described. This experiment was approved by the Ethical Review Committee of Inner-City Faculties at Maastricht University (ERCIC).

Figure 39: Informed consent in the estimation experiment

Welcome to this study!

Welcome | **Your Data** | **Informed Consent**

Your consent to participate

I hereby give permission for my personal data to be used for this research project. I have had enough time to decide whether I want to participate in the experiment. I know that participation is voluntary, and I know that I can decide to withdraw from the experiment at any time. I do not have to justify such a decision to withdraw. If I withdraw, I will forfeit any payments.

I understand that personal data collected throughout the experiment will be stored on secure servers of Maastricht University and only be available to the research team as outlined before. Only fully anonymised data may be made available outside the research team. Moreover, I give permission for the researcher to use my anonymised responses in subsequent experiments.

Participate in this experiment

F.2.2 Introduction to the experiment

Figure 40: Introduction to the decision task

Notes: The introduction video on the decision task can be found in the replication package.

Figure 41: Check of understanding

Notes: Upon hitting the "Check answers" button, participants are informed which questions still contain incorrect answers. They can revisit the introduction video and try to re-answer the questions. Only, after all questions are answered correctly participants may continue in the experiment. The number of attempts is recorded.

Figure 42: Prompt to ask any question that may have remained unanswered after the introduction

F.2.3   Decision task

Figure 43: Decision task screen for evaluating group judgments from FTF groups without hidden agendas

Figure 44: Decision task screen for evaluating group judgments from FTF groups with hidden agendas

Evaluation Phase

FTF groups with hidden agendas

Please evaluate judgment 1 out of 10 from face-to-face groups with hidden agendas

Remember: In face-to-face groups with hidden agendas ...

Two out of four group members have a hidden agenda to **manipulate the judgment either towards 0 or towards 100**. Whether it was one or the other direction was determined randomly for each judgment task the group faced. Both group members had the same hidden agenda. The more they managed to achieve their hidden agenda the higher their chance to receive an **additional bonus**.

Apart from that group interaction is the same as in face-to-face groups without hidden agendas ...

- group members **discussed freely** in a **zoom video call**
- the final group judgment was a **reported consensus** of all group members
- the **more accurate** the group's judgment the higher their chance to receive a **bonus**

Please use the handles on the slider below to indicate the range you think the true value falls into.

group judgment
(face-to-face + hidden agendas)
17.5

lower bound   upper bound
0   100

Figure 45: Decision task screen for evaluating group judgments from Delphi groups without hidden agendas

Figure 46: Decision task screen for evaluating group judgments from Delphi groups with hidden agendas

Evaluation Phase

Delphi groups with hidden agendas

Please evaluate judgment 1 out of 10 from Delphi groups with hidden agendas

Remember: In Delphi groups with hidden agendas ...

Two out of four group members have a hidden agenda to **manipulate the judgment either towards 0 or towards 100**. Whether it was one or the other direction was determined randomly for judgment each task the group faced. Both group members had the same hidden agenda. The more they managed to achieve their hidden agenda the higher their chance to receive an **additional bonus**.

Apart from that group interaction is the same as in Delphi groups without hidden agendas ...

- group members gave a first **individual judgment and reasoning**
- then they **saw the reports of their pseudonymized group members** and gave a second individual judgment
- the final group judgment was the **average of second individual judgments**
- the **more accurate** the group's judgment the higher their chance to receive a **bonus**

Please use the handles on the slider below to indicate the range you think the true value falls into.

group judgment
(Delphi + hidden agendas)

83.2

lower bound   upper bound
0   100

F.2.4 Closing survey

Figure 47: Closing survey - personal characteristics

Figure 48: Closing survey - experience during experiment

Questionnaire

Now, please read the following statement and state how well it describes you as a person.

As long as I am not convinced otherwise, I assume that people have only the best intentions.

does not describe me at all ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ describes me perfectly

Finally, please answer some questions about the task you worked on throughout today's experiment.

How would you rate your own understanding of the task?

very weak ○ ○ ○ ○ ○ very good

Please, briefly describe how you tried to solve the task in the experiment:
How did you come up with a range around the group judgments, based on the information your were given?

```
I used my information in the following way ...
```

Did you evaluate group judgments from groups with hidden agendas differently? How?

```
When hidden agendas were present ...
```

Finally, if you wanted to get the judgment of a group of people, of which some have a hidden agenda, how would you like that group to interact?

```
It would be great if ...
```

Once you answered all questions, please click the blue button in order to continue to the final page, where you will see your earnings from this experiment.

Continue