Mitigating manipulation in committees - the virtue of face-to-face communication

David Albrecht*

August 24, 2023

WORK IN PROGRESS - PLEASE DO NOT CIRCULATE

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 induce a consequent decision in their interest.

Utilizing an incentivized experiment, I analyze how manipulation affects the accuracy and trustworthiness of such group judgment. To this end, I compare group judgments from the scientifically endorsed, structured Delphi interaction as opposed to unstructured face-to-face interaction, which is ubiquitous in real-world institutions. Furthermore, I identify mechanisms of robustness, by decomposing accuracy differences across both techniques into bias, noise, and utilization of valuable information through structural estimations in the BIN-modelling framework (Satopää et al., 2021).

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. As a consequence, with manipulation Delphi is less accurate than face-to-face interaction and even worse than a naïve benchmark. Trustworthiness does not always match 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. Manipulation likely decreases accuracy through more bias and less utilization of valuable information.

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

*Maastricht University (david.albrecht@pm.me). I am thankful to Alexander Brüggen, Thomas Meissner, Martin Strobel, Uri Gneezy, Joshua Becker, Ville Satopää, Marie Claire Villeval, my PhD colleagues at Maastricht University and others for many helpful discussions. Numerous valuable comments were also received from participants in seminars at Maastricht University, University of California San Diego, and at conferences in Maastricht, Nice, and Santa Barbara. Angela Thissen and Kaiqi Liu provided excellent research assistance for analysing the video recordings of the experiment. I acknowledge funding through the GSBE grant for primary data collection, as well as by Alexander Brüggen and Martin Strobel. Replication files can be found via OSF: https://osf.io/wh7fr/?view_only=71c7f9270b7c40608fd6f4e1ff419a6f.

1 Introduction

In many institutions, high-stake decisions are made based on the collective judgment of small groups, like (expert) committees, boards, or teams who evaluate some decision-relevant criterion. Examples of such collective intelligence span diverse contexts: Politicians judge effectiveness of measures to fight global pandemics (Haug et al., 2020), executive boards rate investment alternatives (Lovallo et al., 2020) and security experts assess terrorist activity (Friedman and Zeckhauser, 2014).

Does the widespread use of this practice imply its appropriateness? To optimally inform decisions, the accuracy and trustworthiness of group judgments are key. However, some group members may have a hidden agenda, and manipulate this judgment in a certain direction, to induce a consequent decision in their own interest.[1] Generating accurate and trusted collective intelligence in such a setting remains a challenge. I analyze whether two common interaction formats to elicit a joint group judgment: face-to-face meetings, and the Delphi technique, may preserve accuracy and trustworthiness despite manipulation due to hidden agendas.

To date, a way to extract accurate and simultaneously trusted collective intelligence from groups in hidden agenda settings is 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 response to manipulation, decision-makers observing the market lose trust and make decisions that are even worse as if they would have ignored the mar-

[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), reputational concerns (Visser and Swank, 2007), overstating chances of political victory of favoured parties (Hansen et al., 2004; Rothschild and Sethi, 2016), receiving compensating bribes (Felgenhauer and Grüner, 2008), or downplaying security threat to reduce responsibility and effort for consequent duties (Amjahid et al., 2017).

ket completely (Kaplan et al., 2022). Face-to-face interactions are a standard operating procedure in institutional decision-making. A particular variety of such interactions are nominal groups. These groups interact 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). As a consequence, many objectively ill-informed and thus poor decisions with high societal stakes might be taken. 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 favored Delphi technique (Delphi), where interaction is exogenously structured according to a specific protocol.[3] I quantify to what extent these interaction formats produce collective intelligence that is simultaneously accurate and trusted. Further, I identify what drives robustness towards manipulation.

To measure the capabilities of Delphi and FTF I develop a preregistered, incentivized lab experiment.[4,5] The experiment enables the isolation of the causal effects of hidden agendas and the format of group interaction on quantitative measures of accuracy and trust. 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 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 subse-

[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). The Delphi technique, though less common in practice, is a group interaction format supported by empirical and theoretical research (Rowe and Wright, 1999).

[4]Preregistration 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).

quent *decision experiment*. In the first part, the *estimation experiment*, I examine the 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 treatment arm, groups interact via face-to-face conversation in a video call without any imposed structure. In the Delphi treatment arm groups interact mediated through an imposed, pseudonymous chat protocol. According to this protocol, group members first produce a quantitative judgment and qualitative reasoning individually, then observe the individual pseudonymous judgments and reasonings of all group members, and finally produce a second quantitative estimate individually. Across all treatment arms, I incentivize accuracy at the group level. In the hidden agenda treatment arm, I additionally induce manipulation incentives (hidden agendas) in half of the group members through individual, and private side payments.

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

Comparing absolute errors across conditions of the estimation experiment, I identify differences in accuracy. 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 set accuracy into perspective, I compare Delphi and FTF groups against the best Bayesian benchmark based on the information available to groups. Without hidden agendas, best benchmark's accuracy is statistically indistinguishable from Delphi's accuracy but significantly more accurate than FTF. Both interaction formats, Delphi and FTF, perform significantly better than a negative benchmark, which estimates all probabilities naïvely at 50%. In situations with hidden agenda, no interaction format reaches the best Bayesian benchmark, FTF groups

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

outperform the naïve heuristic, and Delphi falls behind even this negative benchmark.

Comparing the length of stated confidence intervals in the decision experiment, across conditions of the estimation experiment, I identify differences in the trustworthiness of group judgments. 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 accuracy. As a consequence, FTF group judgments appear problematic in situations without hidden agendas. They are less accurate, decision makers fail to realize this, and assign relatively more trust to FTF judgments. This may be taken as evidence that FTF should be used less in institutional decision making in situations without hidden agendas. The Delphi technique seems to be a better alternative. However, in situations with hidden agendas trustworthiness is aligned with accuracy, as the relatively more accurate judgments from FTF groups are also trusted more than judgments from Delphi groups. This may be taken evidence that FTF might be an appropriate format for use in institutional decision making in situations where hidden agendas might play a role.

Further analysis on the estimation experiment, is exploratory and seeks to disentangle the sources of accuracy differences across conditions. For this purpose, the data from group estimation is analyzed in the BIN-modelling framework (Satopää et al., 2021) (Section 4.3.1). This allows 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. As such, the Delphi technique appears preferable in situations without hidden agendas but FTF seems better suited for situations with hidden agendas.

2 Related literature

This work connects to the wide 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. Farther, 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 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. Importantly, many studies used (over)simplify Delphi designs, which in fact might have led to Delphi's capacity being understated (Rowe et al., 1991). Later, also prediction markets 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, the judgment of nominal groups are perceived as more trustworthy overall and 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 completely. I summarize the designs of technique comparisons most relevant to the present research in Table 1.

To date it remains a challenge to extract collective intelligence from groups in hidden agenda settings that is simultaneously accurate and trusted. As such, 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 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 have an effect on other participants, and to ultimately influence decisions, which are tied to the Delphi result. Yet, explicit manipulation of Delphi estimates by group members has not been studied empirically before this investigation.

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 judge (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, but 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 the implementation of 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 kind of voting procedure, which may be preceded by deliberation. This has common features with the given setting of groups deriving a joint probabilistic judgment, but also exhibits clear differences, as e.g. the given setting does not explicitly study certain 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). Making such deliberation public to non-committee members however counteracts these positive effects (Fehrler and Hughes, 2018). Moreover, if committee members differ in competence, transparency of the voting behavior may counteract hidden agendas of competent members (Mattozzi and Nakaguma, 2022).

Furthermore, effects of manipulation on collective intelligence strongly relate to the general literature on **lying and deception**. In contrast to predictions of standard economic theory, a substantial amount of individuals tells 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).

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)
nominal group technique	✓(3-6)	✓(3)	✓(3)	-	-
Delphi technique	✓(4-6)	-	-	-	✓(4)
prediction markets	✓(4-7)	✓(3)	✓(3)	✓(8)	-
Judgment task					
content	factual questions	hidden profiles / Wason task	hidden profiles	hidden profiles	probability judgment
induced expertise	-	✓	✓	✓	✓
Incentives					
accuracy (excl. markets)	discrete tournament	discrete scoring rule	discrete scoring rule	-	proper scoring rule
hidden agendas	-	discrete scoring rule	discrete scoring rule	discrete scoring rule	proper scoring rule
trust	-	-	-	discrete scoring rule	most likely interval method
Measure of assessment					
accuracy	absolute errors (AE)	# of correct solutions / market prices	# of correct solutions	market prices	AE / Brier scores
trust	-	-	Likert scale	decisions	confidence intervals
Sample					
n judgment	227	144	450	112	590
groups/ condition	11	16	15	7	15
n trust	-	-	224 + 358	56	2000

3 Experiment

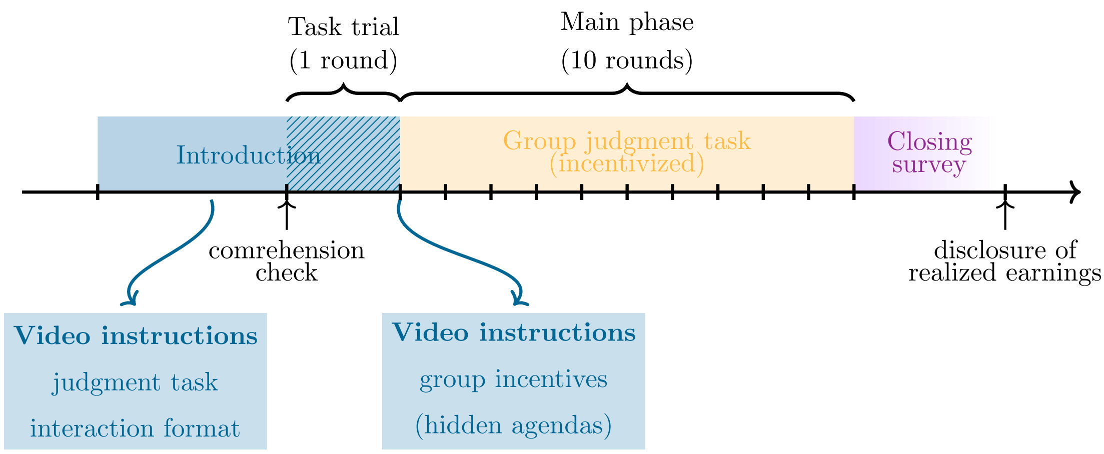
The study comprises two complementary experiments: the *estimation experiment*, to examine the accuracy of group judgments, and the subsequent *decision experiment*, to elicit the trustworthiness of group judgments. All experiments are programmed in oTree (Chen et al., 2016) and were implemented onsite in the experimental economics laboratory BEElab at Maastricht University in 2022 and 2023. 240 people participated in the estimation experiment, split up into 60 groups of four, 15 groups per treatment condition. Further 50 participants, subsequently, 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 clear mode of the educational backgrounds is in the area of business and/or economics (79.6% in the estimation experiment and 74% of participants in the decision experiment). A majority of 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 treatment arm, 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).

In all four conditions the estimation experiment follows the same structure, 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 the judgment task and the interaction format. After the trial, before starting the

Figure 1: Timeline of the estimation experiment



main phase, participants are informed about the incentive scheme, and consequently also whether they interact in a setting with hidden agendas. In settings with hidden agendas, participants get to know their own type (with or without hidden agenda), and that exactly two group members have a hidden agenda and the remaining two group members have no hidden agenda. Thus, knowledge about the own type allows to infer the distribution of types among the remaining group members. 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 agenda in any given round. In a specific round, the two group members with hidden agenda know that their hidden agenda in this round is to drive estimates up (down), but this information is not disclosed to the remaining two group members without hidden agenda. Types do not change throughout the whole experiment. 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 members 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 very end of the experiment. Before earnings disclosure and payment, participants complete a questionnaire on socio-demographic characteristics as well as 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 that is 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** treatment arm. 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 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 engage in a direct discussion but rather interact through 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 the inputs of the three remaining group members, and their own input again. This feedback is shown pseudonymous as "the input of Person A, B, and C", and "your input". In this way, it is not possible to link the estimates and reasonings to a specific real person. 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 the group judgment, as well as the underlying, second, individual estimates before proceeding in 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 comprise ten steps $X_t = X_1, X_2, ..., X_{10}$. Each step X_t may take the values -1, 0 and 1. X_t is determined by independent draws from an identical distribution. 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. Over the course of 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. To participants, the task is presented 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 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 to evaluate the 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 are introduced to the judgment task, and are informed about the details of interaction format and incentives in each of the four treatment conditions in the estimation experiment. Subsequently, for each estimate, decision-makers state the lower and the upper bound of an interval, which they think likely contains the true value which has been estimated in the estimation experiment. The width of stated confidence

[9]This is an adaption to the preregistration, that allows all estimates of the estimation experiment to be evaluated most balanced: each estimate is evaluated at least three times and by three different decision-makers. The preregistration 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.

intervals can be seen as a measure of the 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 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. In particular, I test three preregistered 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.

Based on the estimation experiment Delphi groups appear more accurate than FTF without hidden agendas, supporting H1. Hidden agendas impair accuracy for Delphi groups, partly in line with H2. However, hidden agendas do not affect accuracy of FTF groups, contradicting H3.

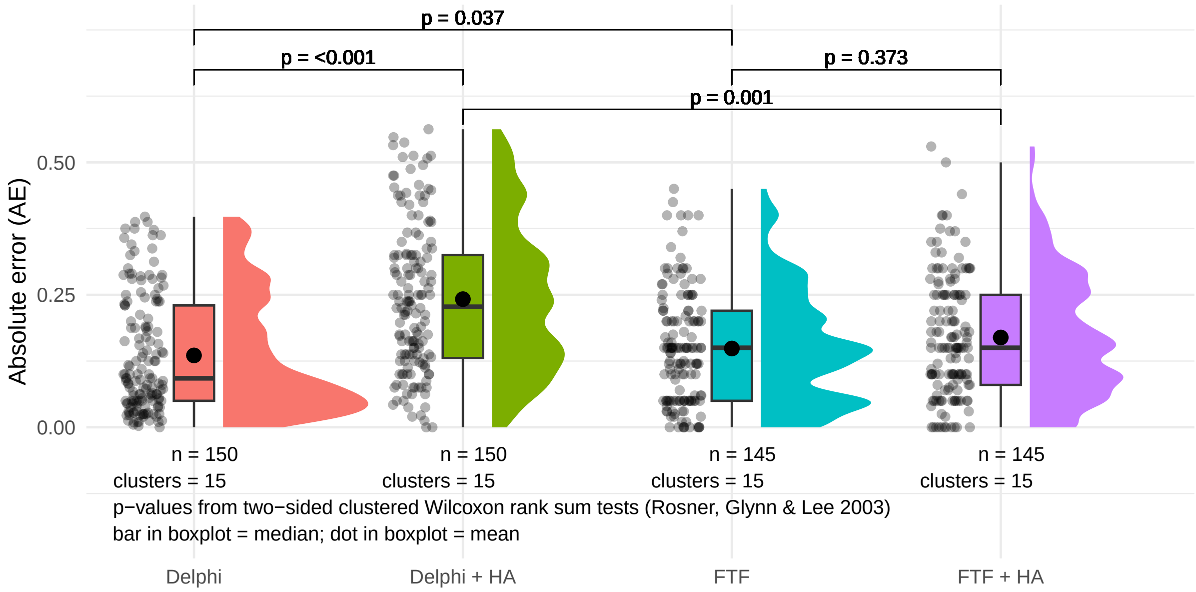
For the main analysis, I quantify accuracy by juxtaposing group judgments to objectively true probabilities.[12] I focus on absolute error (AE)[13] as a simple, and illustrative

[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.

[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 to compare a group judgment to the actual outcomes drawn according to the underlying true 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).

Figure 2: Absolute errors of group judgments across conditions of the estimation experiment



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 stated 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]

[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, which is 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.

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

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. In fact, 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 interaction.*

4.1.1 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) and on the negative side I consider first the naïve heuristic of always stating a “50/50” (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 both 50/50 and uniform, while Delphi falls is less accurate than 50/50.

Figure 3: Absolute errors of probability judgments from FTF groups against best and worst benchmarks

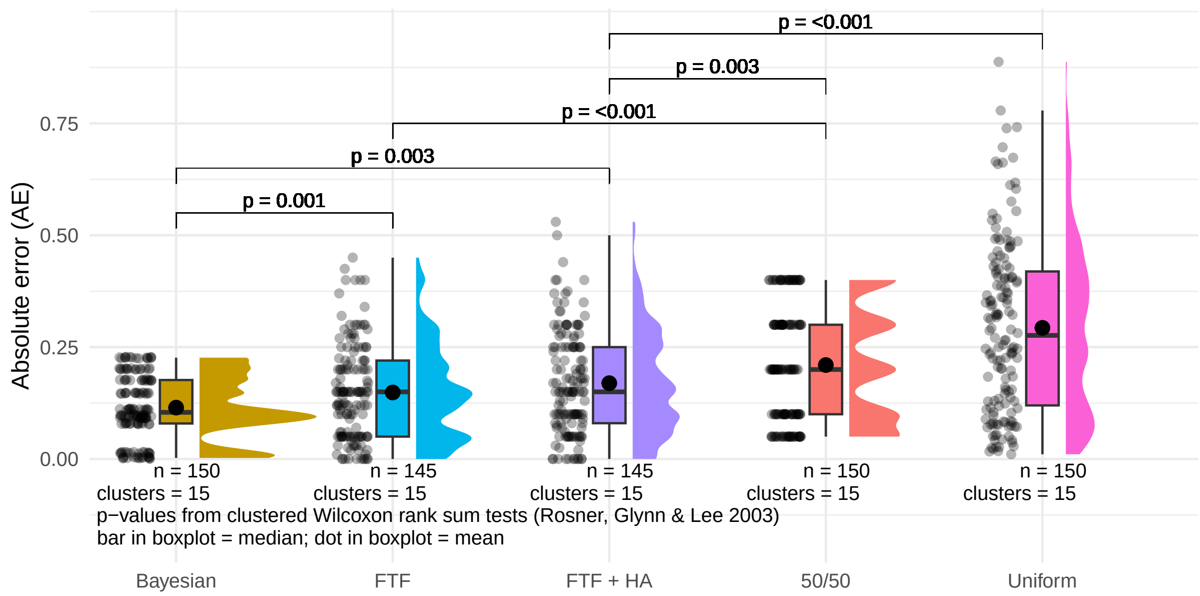
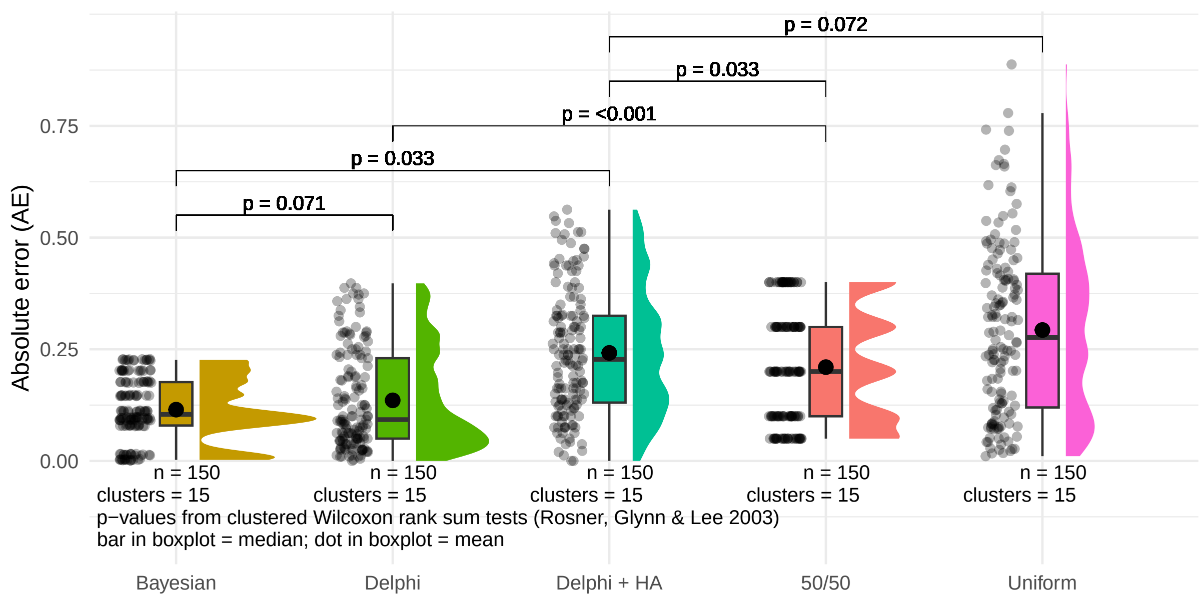


Figure 4: Absolute errors of probability judgments from Delphi groups against best and worst benchmarks



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 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 always 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 both, 50/50 and Uniform, even with hidden agendas.*

To summarize the findings on 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 the 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).

4.1.2 Accuracy and hidden agenda achievement over time

Previously presented results, focus on the accuracy of group judgments, irrespective of whether the judgment result 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? A potential explanation for improving accuracy is that groups may learn to interact better, and improve their understanding of the task. In the same way, in groups with hidden agendas, group members with hidden agenda may learn how to better achieve their hidden agenda over time, and thus impair 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 agenda achieve their hidden agenda.

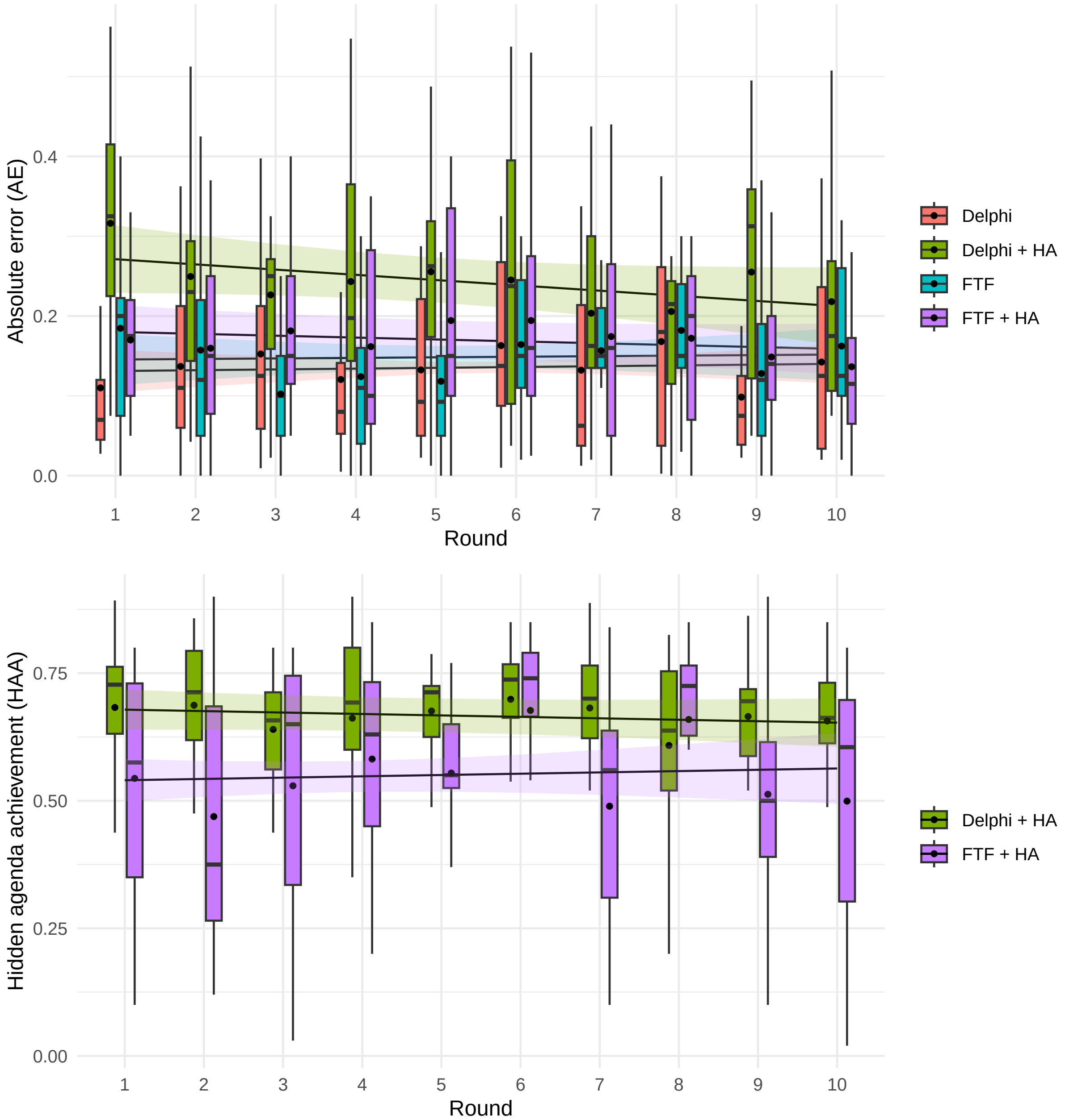
Separating group judgments for the consecutive rounds 1,2, ..., 10 of the judgment task, I visualize accuracy and hidden agenda achievement over time in Figure 5. Based on a general overview, in most rounds the distribution of absolute errors tends towards higher absolute errors for Delphi groups with 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 is in line with the general results on accuracy aggregated over all rounds as depicted earlier in Figure 2. To identify potential trends over time, I regress the absolute error as dependent variable on the round as 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 agenda achieve their hidden agenda 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 full, 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

Figure 5: Accuracy of group judgments and hidden agenda achievement over time



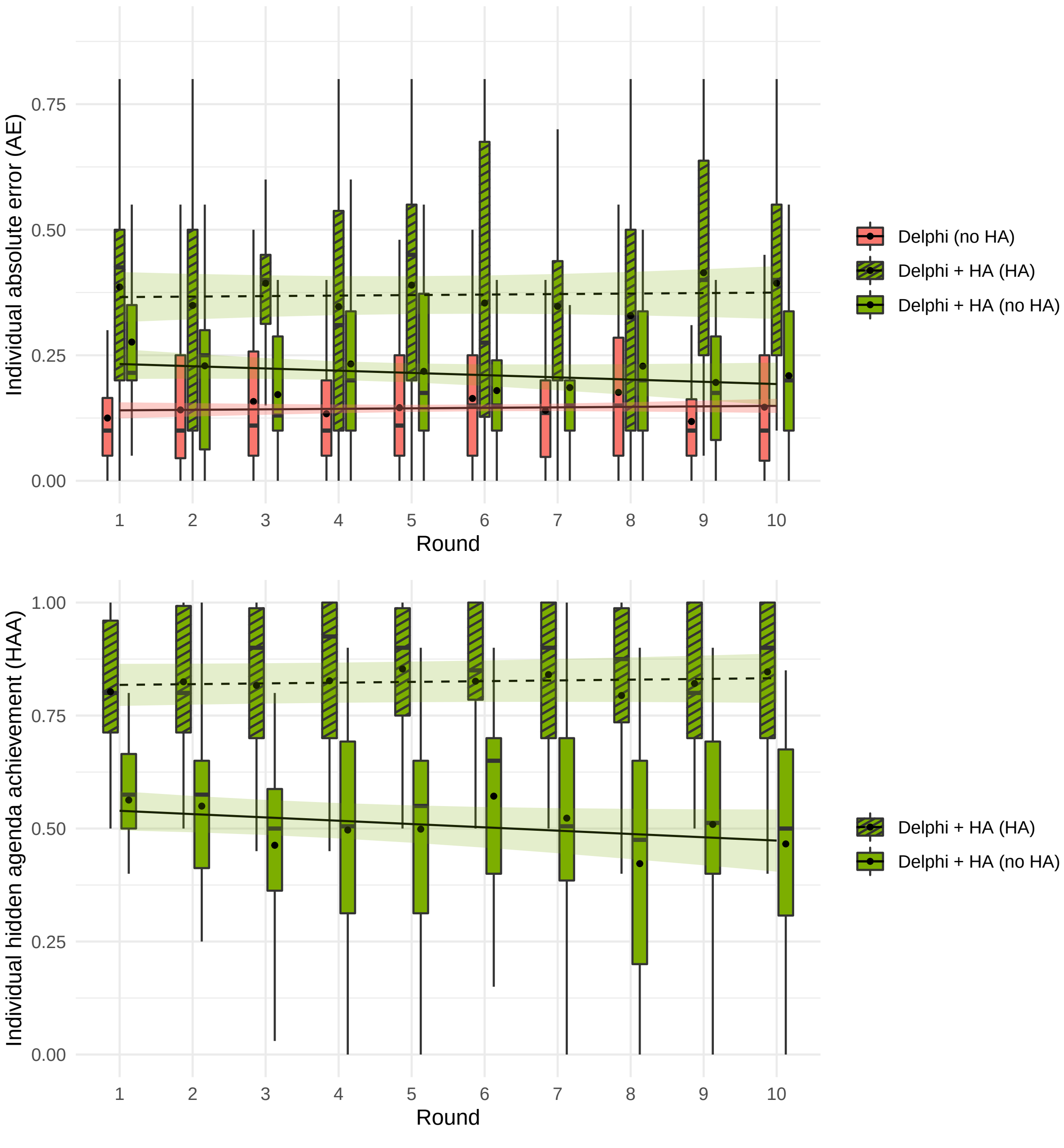
Notes: Whiskers in the boxplots span observations larger than $Q1 - 1.5 * IQR$ and smaller than $Q3 + 1.5 * IQR$. Regressions are OLS with standard errors clustered at the group level.

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. To incorporate the time dimension, I regress the hidden agenda achievement rate on the round. This reveals, just as for accuracy, no statistically significant association between these variables. Again 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.

For Delphi groups, I may further extend the analysis to the individual level. In any Delphi interaction each group member provided a second individual judgment after observing the first individual judgments and reasonings of their group members. The group judgment, was then calculated as the mean of all second individual judgments. Just 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 agenda may over time distort their individual judgments more and more into the direction of the hidden agenda, as they care less about loosing credibility in the decreasing number of future interaction with their group. Alternatively, they might distort less over time, as they become more cautious to not repeatedly behave in an an overly suspicious way. 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 some evidence for such counteracting behaviour. In Delphi groups with hidden agendas, those group members who do not have a hidden agenda themselves, over time shift their second individual judgments more and more in the opposite direction of the hidden agenda.

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 errors for group members with hidden agenda in Delphi groups with hidden agendas, towards lower absolute errors for group members in Delphi groups without hidden agendas, and group members without hidden agenda in Delphi groups with hidden agendas are situated in between. This suggests, that group members with hidden agenda do not only negatively influence the accuracy of Delphi group judgments, but also the second individual judgments of members without hidden agenda, in those groups. Further, in Delphi groups with hidden agendas, individual judgments of group members with hidden agenda are, plausibly, shifted more in the direction of the hidden agenda than those of group members without hidden agenda. Focusing on the time dimension, here I regress

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.

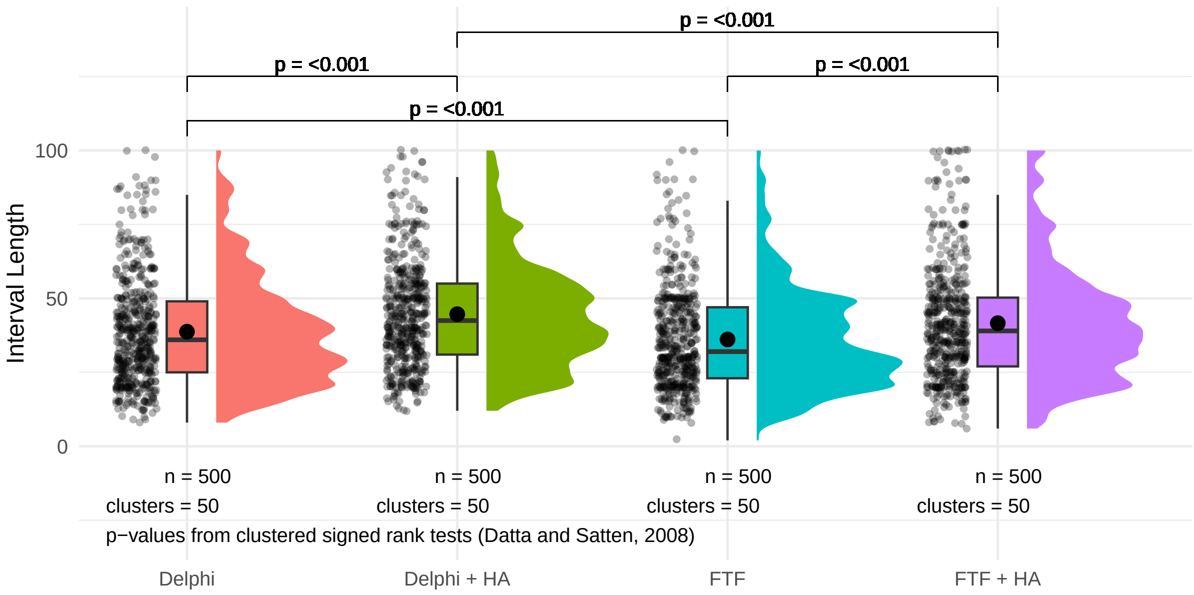
the individual level absolute error and hidden agenda achievement rate on rounds. This reveals, that both individual accuracy and hidden agenda achievement are mostly stable over rounds. Almost all regression lines have a slope that is statistically indistinguishable from zero. The only exception appears in Delphi groups with hidden agendas. For those group members who do not have a hidden agenda themselves, hidden agenda achievement is negatively associated with rounds with marginal significance ($P = 0.0899$).

4.2 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. In particular, I test three preregistered 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. 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 the similar magnitude for FTF and Delphi, supporting H6.

To quantify the trustworthiness of group judgments, I consider the width of confidence intervals. The narrower the confidence interval the more trustworthy the group 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 exactly equal to the true 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

Figure 7: Width of confidence intervals around group judgments from different treatment conditions in the estimation experiment



rank tests adapted for clustered paired data (Datta and Satten, 2008; Jiang et al., 2020).[16]

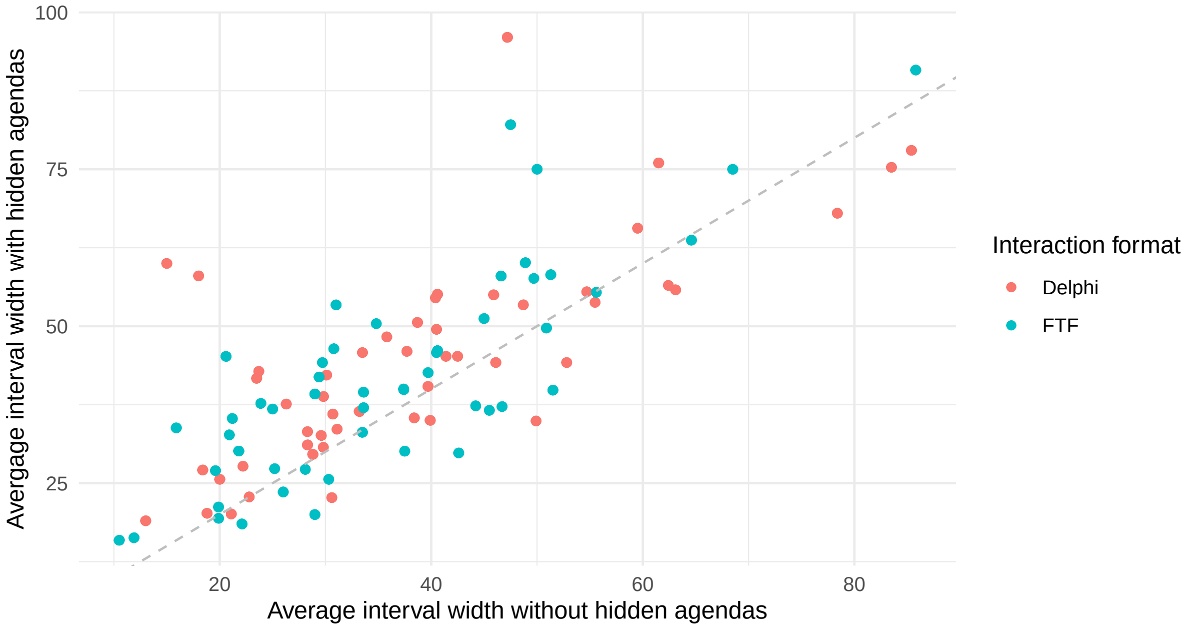
Result 4: *In contrast to H4, FTF group judgments enjoy significantly higher ($P < 0.001$) levels of trust, i.e. narrower stated confidence intervals, than Delphi group judgments. This holds with and without hidden agendas.*

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

[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 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



measures the difference in trustworthiness with and without hidden agendas. As such, points above the 45-degree line correspond to a 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 of them are further away from the line, than points that lie below the 45-dergee line. Further, if the impact of hidden agendas on trustworthiness is similar for FTF and Delphi, we should see no clusters of points representing FTF and Delphi group judgments respectively, e.g. Delphi points generally lying 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.

4.3 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 accuracy judgments: 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 consistently produce too high or too low estimates, noise leads to non-systematic variability in estimates that is not related to the event of interest, and finally, information characterizes the group judgments' variability which is correlated with the event of interest. Bias and noise are detrimental, while information 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, as opposed to 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, which is subject to three latent variables of group accuracy: bias, information 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 difference in the means of Z^* and Z_i, or noise as variability of Z^* that is uncorrelated with Z_i. Accuracy-enhancing information 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:

$$\begin{aligned} Outcome: &\quad Y = \mathbf{1}(Z^* > 0) && (4) \\ Bias: &\quad \mu_i = \mathbb{E}[Z_i] - \mathbb{E}[Z^*] && (5) \\ Information: &\quad \gamma_i = Cov(Z_i, Z^*) && (6) \\ Noise: &\quad \delta_i = Var(Z_i) - Cov(Z_i, Z^*), && (7) \end{aligned}$$

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 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 tried 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. average bias of FTF groups with hidden agendas across the ten judgment tasks in the estimation experiment.

4.3.1 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 ef-

[17]This seems reasonable, as groups as a whole 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.

fects 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 priori 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 then updated according to Bayes rule, taking observed group judgments, i.e. the data, into account. I modify the BIN-tools 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 expected the magnitude of changes in accuracy that can be attributed to each mechanism?

In terms of the likelihoods 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 is a conservative assumption on the prior distribution. The 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 is 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: $\mathbb{P}(\lvert\mu_1\rvert < \lvert\mu_0\rvert)$	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: $\lvert\mu_0\rvert - \lvert\mu_1\rvert$	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: $\delta_1 - \delta_0$	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, while 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 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 ($\mathbb{P} = 0.842$).

For Delphi groups with hidden agendas, an increase in bias is about as likely ($\mathbb{P} = 0.851$) 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 as compared to situations without hidden agendas. 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 level of significance used when interpreting p-values in frequentist settings. Statements on the likelihood of information and bias driving accuracy differences between situations with and without hidden agendas should thus be interpreted with caution.

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 drives a decrease in 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 ($\gamma_0 - \gamma_1$), 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 ($\delta_1 - \delta_0 = 0.63$) and with hidden agendas ($\delta_1 - \delta_0 = 0.66$). The respective 95% credible intervals around the estimates of $\delta_1 - \delta_0$ 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 of the three mechanisms 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) make use of the Good Judgment Project data which comprises between 87 and 191 forecasts per person, the estimation experiment only yields ten distinct judgments per group.[21]

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 as well as 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

[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. On the flip-side 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.

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 of 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 accuracy-impairing mechanism, that may explain differences between FTF and Delphi. While Delphi group judgments likely suffer from increased bias in 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 goes in favor of 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 on 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, also modifications of group interaction formats that vary by one detail at a time may be tested against each other. Beyond this, the investigation on 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 on 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.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–1153. Cited on page 9.

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 8.

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 8.

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 8.

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 8.

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 8.

Bond, C. F., J. and B. M. DePaulo (2006): "Accuracy of deception judgments," *Pers Soc Psychol Rev*, 10, 214–34. Cited on page 9.

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 8.

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 9.

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 8.

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 8.

Dalkey, N. C. (1975): "Toward a theory of group estimation," *The Delphi method: Techniques and applications*, 236–261. Cited on page 8.

Datta, S. and G. A. Satten (2008): "A signed-rank test for clustered data," *Biometrics*, 64, 501–7. Cited on page 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 7.

Fehrler, S. and N. Hughes (2018): "How Transparency Kills Information Aggregation: Theory and Experiment," *American Economic Journal: Microeconomics*, 10, 181–209. Cited on page 9.

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 9.

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 9.

Galton, F. (1907): "Vox populi," *Nature*, 75. Cited on page 8.

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 9.

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 9.

Goeree, J. K. and L. Yariv (2011): "An Experimental Study of Collective Deliberation," *Econometrica*, 79, 893–921. Cited on page 9.

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 6, 10, 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 8 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 8.

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 7.

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 7.

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 9.

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 9.

Hossain, T. and R. Okui (2013): "The Binarized Scoring Rule," *The Review of Economic Studies*, 80, 984–1001. Cited on page 48.

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 and 26.

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 8.

Kaplan, T. R., L. Choo, and R. Zultan (2022): "Manipulation and (Mis)trust in Prediction Markets." *Management Science*. Cited on pages 3, 6, 7, and 10.

Khalmetski, K., B. Rockenbach, and P. Werner (2017): "Evasive lying in strategic communication," *Journal of Public Economics*, 156, 59–72. Cited on page 9.

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 6, 10, 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, 6, 10, 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 9.

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 9.

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 9.

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 46.

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 8 and 35.

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 7.

Parenté, F. and J. Anderson-Parenté (1987): *Delphi inquiry systems*, Chichester: Wiley, 129–156. Cited on page 8.

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 page 17.

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 8.

——— (1999): "The Delphi technique as a forecasting tool: issues and analysis," *International Journal of Forecasting*, 15, 353–375. Cited on pages 3, 6, 8, and 13.

——— (2001): *Expert Opinions in Forecasting: The Role of the Delphi Technique*, Boston: Springer. Cited on pages 7, 8, 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, 8, 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 8.

Satopää, V., M. Salikhov, P. Tetlock, and B. Mellers (2021): "Bias, Information, Noise: The BIN Model of Forecasting," *Management Science*. Cited on pages 1, 5, 28, 30, 34, and 53.

Satopää, V., M. Salikhov, and E. Moreno (2022): "BINtools," . Cited on pages 31 and 53.

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 49.

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 9.

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 9.

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 46.

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 9.

Teschner, F., D. Rothschild, and H. Gimpel (2017): "Manipulation in Conditional Decision Markets," *Group Decision and Negotiation*, 26, 953–971. Cited on page 7.

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 20.

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 8.

WITTROCK, L. (2023): "Useful forecasting: belief elicitation for decision making," *Working Paper*. Cited on page 7.

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, 8, and 14.

A Experimental details

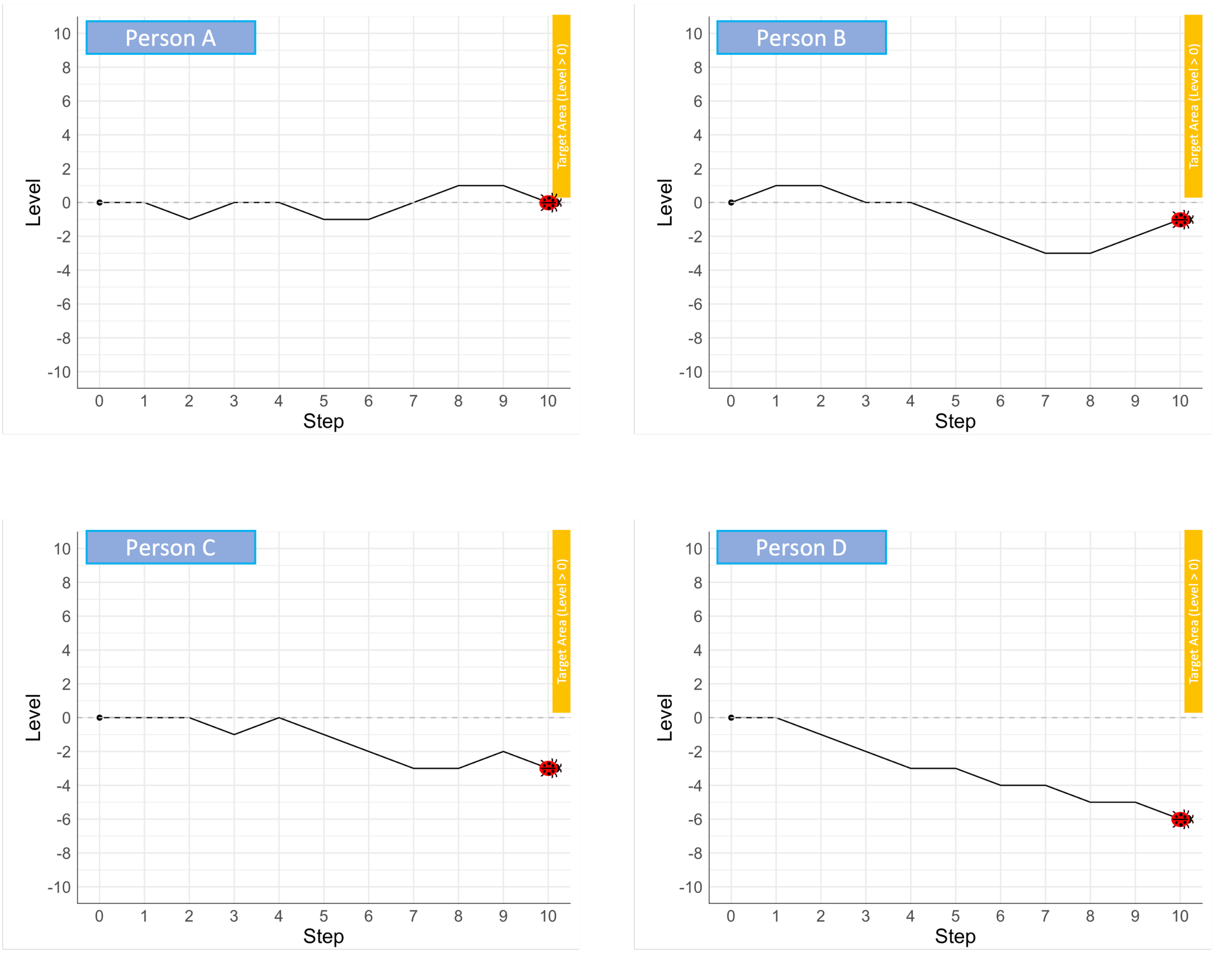
A.1 Estimation experiment

A.1.1 Judgment task

To ease understanding and to increase engagement of experimental participants, the judgment task is framed in a gamified story to 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 terms of frequencies, such as 75 in 100 to facilitate better statistical reasoning (McDowell and Jacobs, 2017). As information each of them sees a movement path generated by a distinct past execution of the program, compare Figure 9. For each round group members are informed that the movement paths are generated by a distinct computer program and consequently the underlying probability of reaching the target might be different.

All in all, 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 full control of 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 but also hidden agenda actions. Third, the behavior of participants can be evaluated against theoretically optimal behavior given the information they have at hand. Finally, probabilistic judgment caters to a wide variety of real-world applications that goes far beyond the motivating examples presented in the introduction.

Figure 9: 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, in the hidden agenda treatment arm, two group members can earn a supplementary individual bonus, by influencing the group judgment in a particular direction.

For each of the 10 rounds of the judgment task, all groups receive an **accuracy-based bonus** which is 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 each individual has the incentive 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 larger the chance that 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 $Y \sim U(0, 1) > p_r^2$.[22] By way of illustration consider the true chance that the movement path reaches the target is $p^* = 0.7$, i.e. 70 in 100 in frequency terms and the group estimated the chance at $p_r = 0.6$, i.e. 60 in 100. 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 receive an **individual hidden agenda bonus** of €1.50 per round also based on a binarized scoring. Their hidden agenda is to drive the group estimate either to 100 in 100 or 0 in 100, i.e. either up or down. Whether it is one or the other direction is determined randomly in each round. The direction of the hidden agenda is

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

always the same for the two group members with a hidden agenda in a given round. The hidden agenda bonus is designed 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 0 in 100 and p_r^2 if the hidden agenda is to drive the estimate 100 in 100.

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:

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

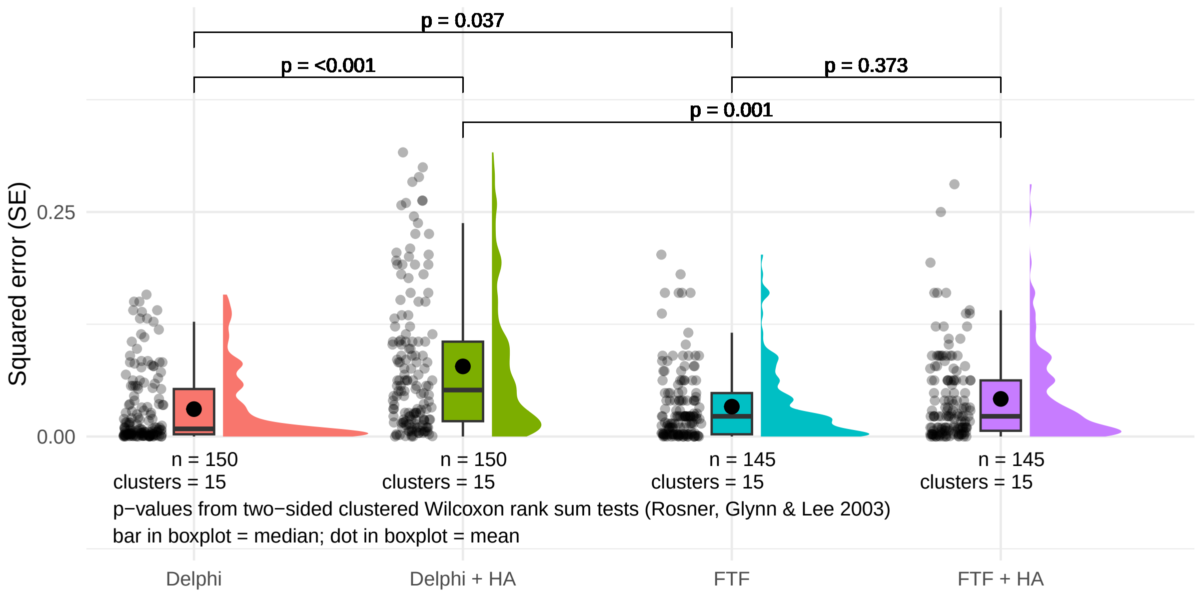
depending on the respective true probability p^*. Note, that estimates generated in the estimation experiment are naturally bounded by 0 and 100. Further, the width of the stated confidence interval ranging from lower bound L to upper bound U is $W = U - L$, which is at most 100. Consequently, the wider 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 10) and Brier Scores (Figure 11). 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 potential dependence of observations.

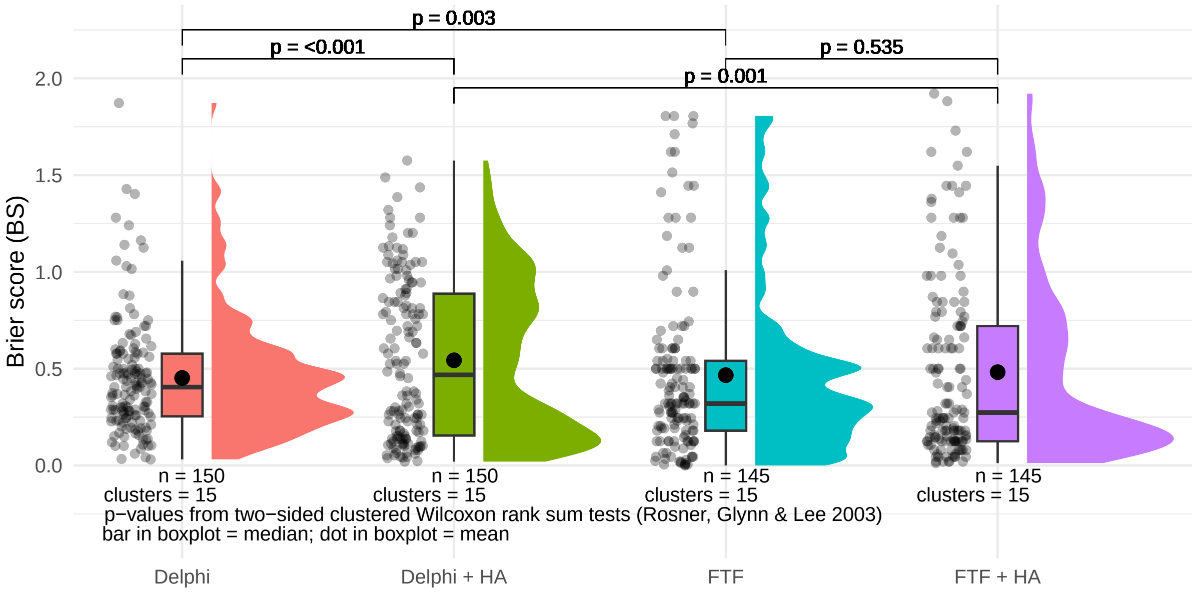
Figure 10: Squared errors of group judgments across experimental conditions



Squared errors (SE), just as 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 11: Brier Scores of group judgments across experimental conditions



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 10 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 dependence of judgments made within the same group, I report results of Wilcoxon rank sum tests for average AE per group (most strict), in Table 3. The most strict scenario posits, that groups make 10 judgments over the 10

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

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. The accuracy of FTF groups and Delphi groups 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 10 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 better accommodate 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 judgement 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. [23] 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

[23]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.

Online Appendix

D Coding scheme for video recordings

D.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)

D.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

E Experimental instructions

E.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.

E.1.1 Welcome information

Figure 12: Opening screen of the estimation experiment

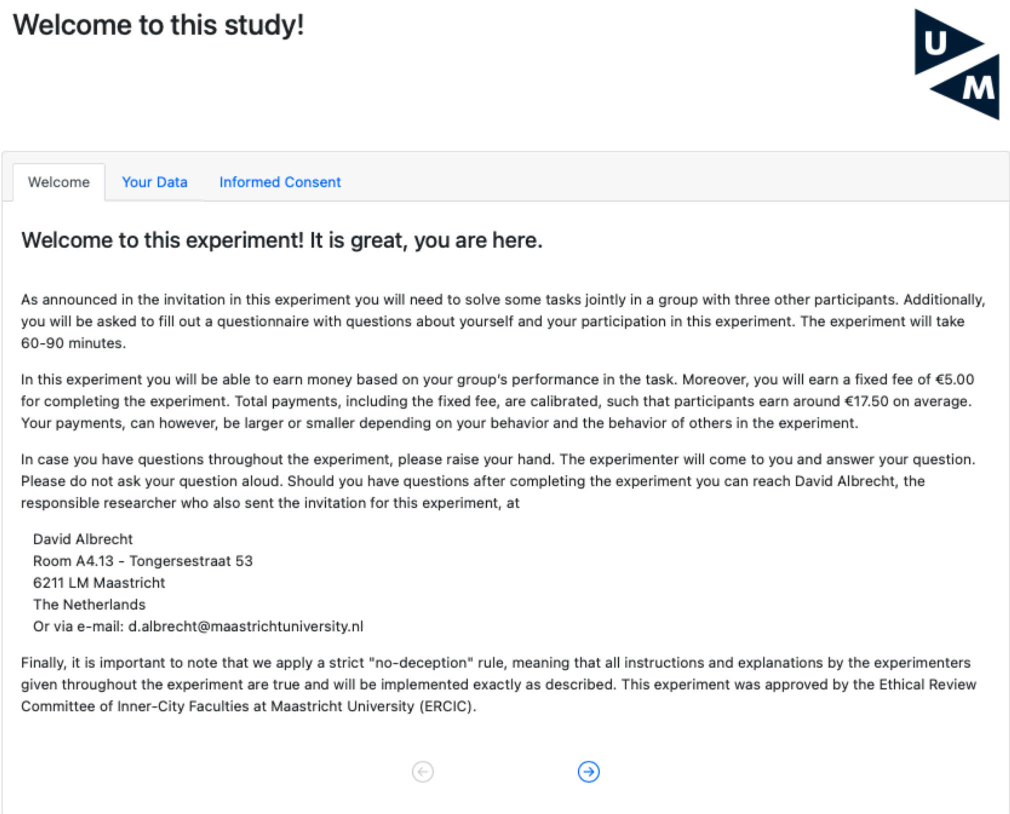


Figure 13: Information on data handling in the estimation experiment

Welcome to this study!

Welcome | Your Data | Informed Consent

Here you can find the details on our data protection policy

Upon your consent, the responses to the experiment you provide will be collected. Additionally, the experiment will be video recorded.

Until completion of the project all data, including video recordings, is accessible to the research team comprising the following employees of SBE: Alexander Brüggen, Martin Strobel, Thomas Meissner and David Albrecht. Furthermore, the research team may grant data access to research assistants sourced from researchers and students at UM upon signing a confidentiality agreement. The videos will be stored for the time span of the underlying study in order to extract anonymous information. Your video recordings will be deleted upon completion of the transcription and extraction of anonymous information.

After completion of the project, anonymized results and anonymized data may be published in academic papers, and data- as well as research repositories. These materials will only comprise anonymous information. It is not possible to link this material to an individual person.

What is the legal basis for holding these data?

The lawful basis for processing this information is your consent which you will be asked to give at the next page.

Your data will not be used for other purposes. Only fully anonymized data will be made available outside the research team. This happens for the sole purpose of replicating the analysis if requested e.g. by scientific journals. You have the right to request access to your personal data and/or deletion by sending an email to d.albrecht@maastrichtuniversity.nl.

How do we store the data?

This experiment uses the app oTree, which is hosted on a local server here at Maastricht University. All raw data collected in the experiments is stored securely on servers at Maastricht University. Personal data, including your video recordings will be encrypted and deleted after completion of transcription and extraction of anonymized information. Maastricht University stores all anonymous research data for at least 10 years. After that, the data is destroyed or transferred to other media for longer storage if needed.

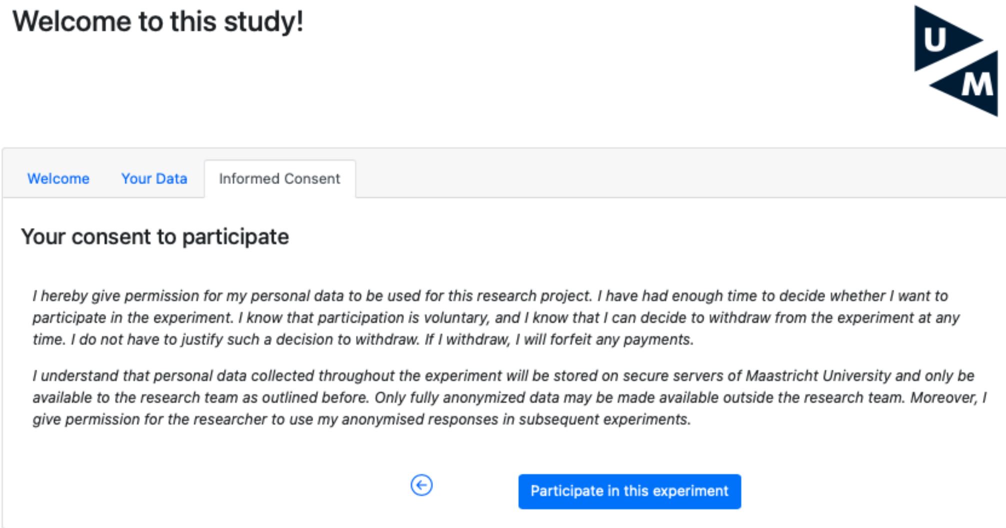
If you have questions, comments or concerns about the data handling of this research project, you can contact the responsible researcher David Albrecht at d.albrecht@maastrichtuniversity.nl.

If you have any specific questions regarding the handling of personal data, you can also submit these to the Data Protection Officer by sending an email to fg@maastrichtuniversity.nl. You also have the right to lodge a complaint with the Dutch Data Protection Authority.

Read less

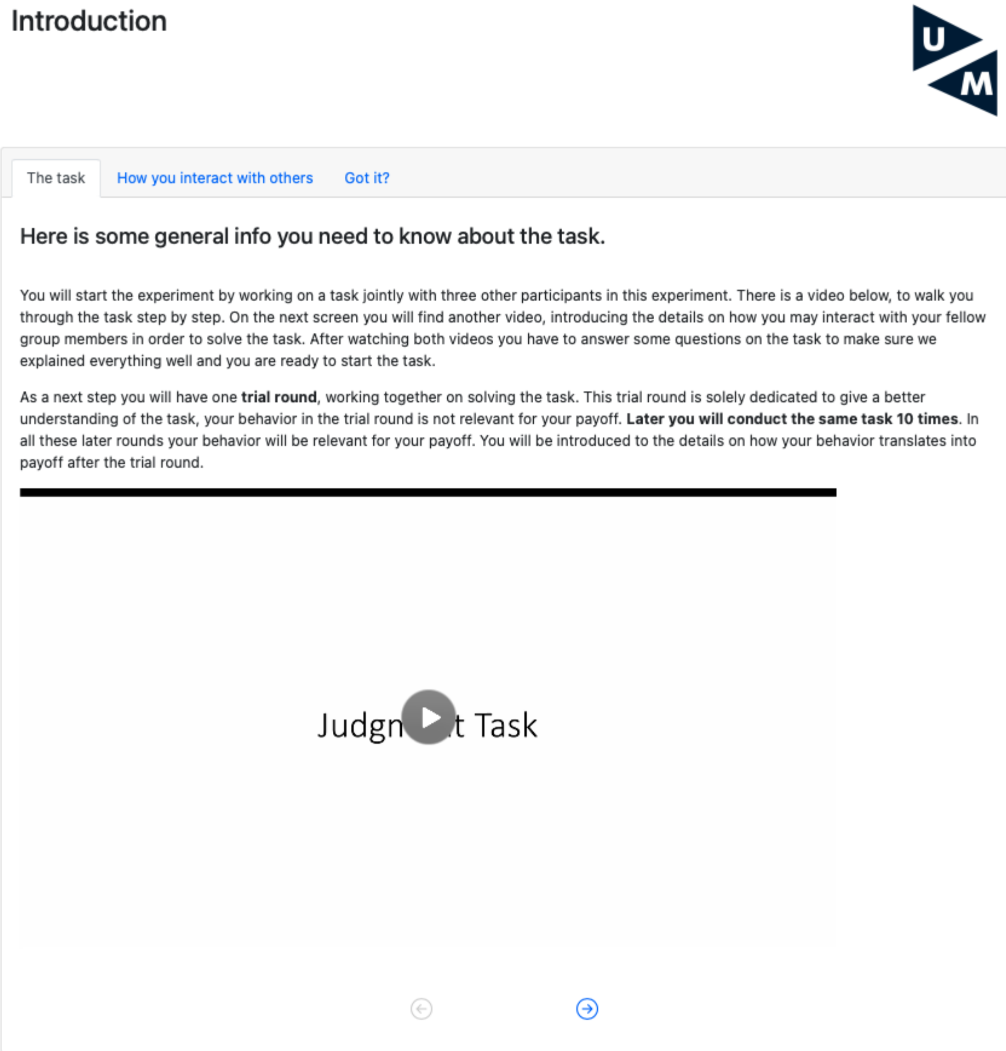
Notes: This screen is part of a face-to-face treatment, and participants are informed that the experiment will be video recorded. In the Delphi treatments there is no video recording and the respective information is not displayed to participants.

Figure 14: Informed consent in the estimation experiment



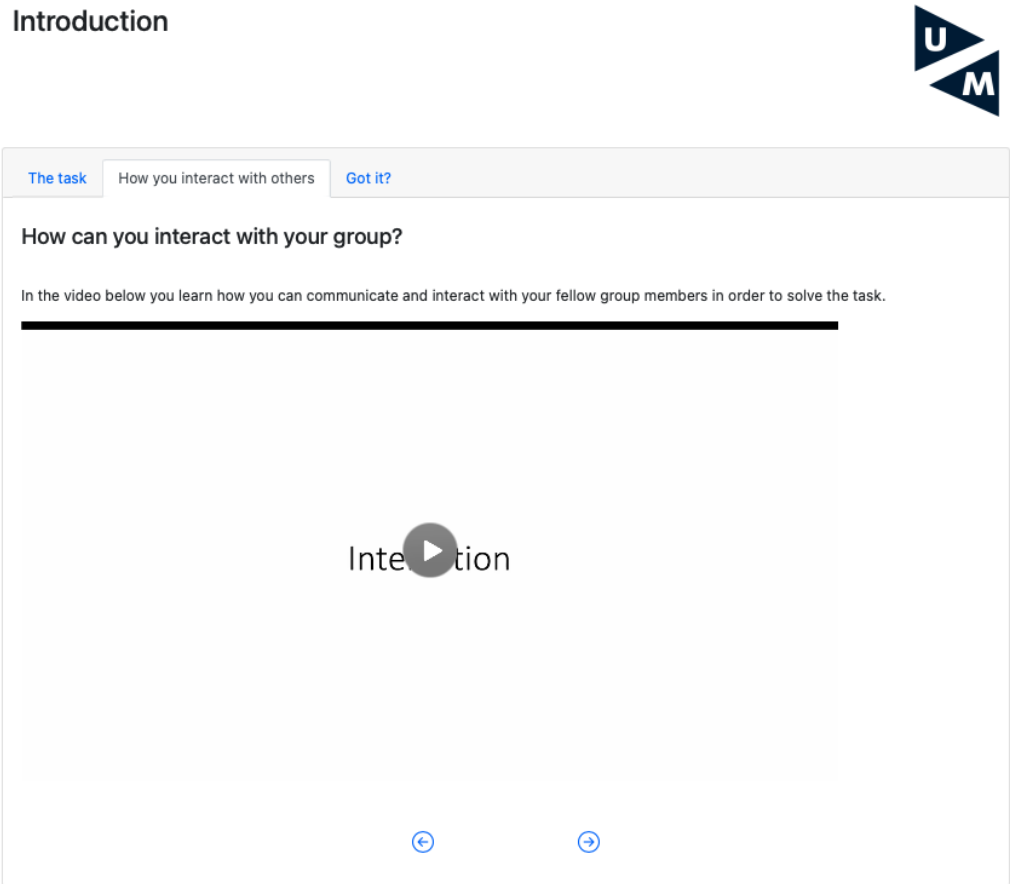
E.1.2 Introduction to the experiment

Figure 15: Introduction to the group judgment task



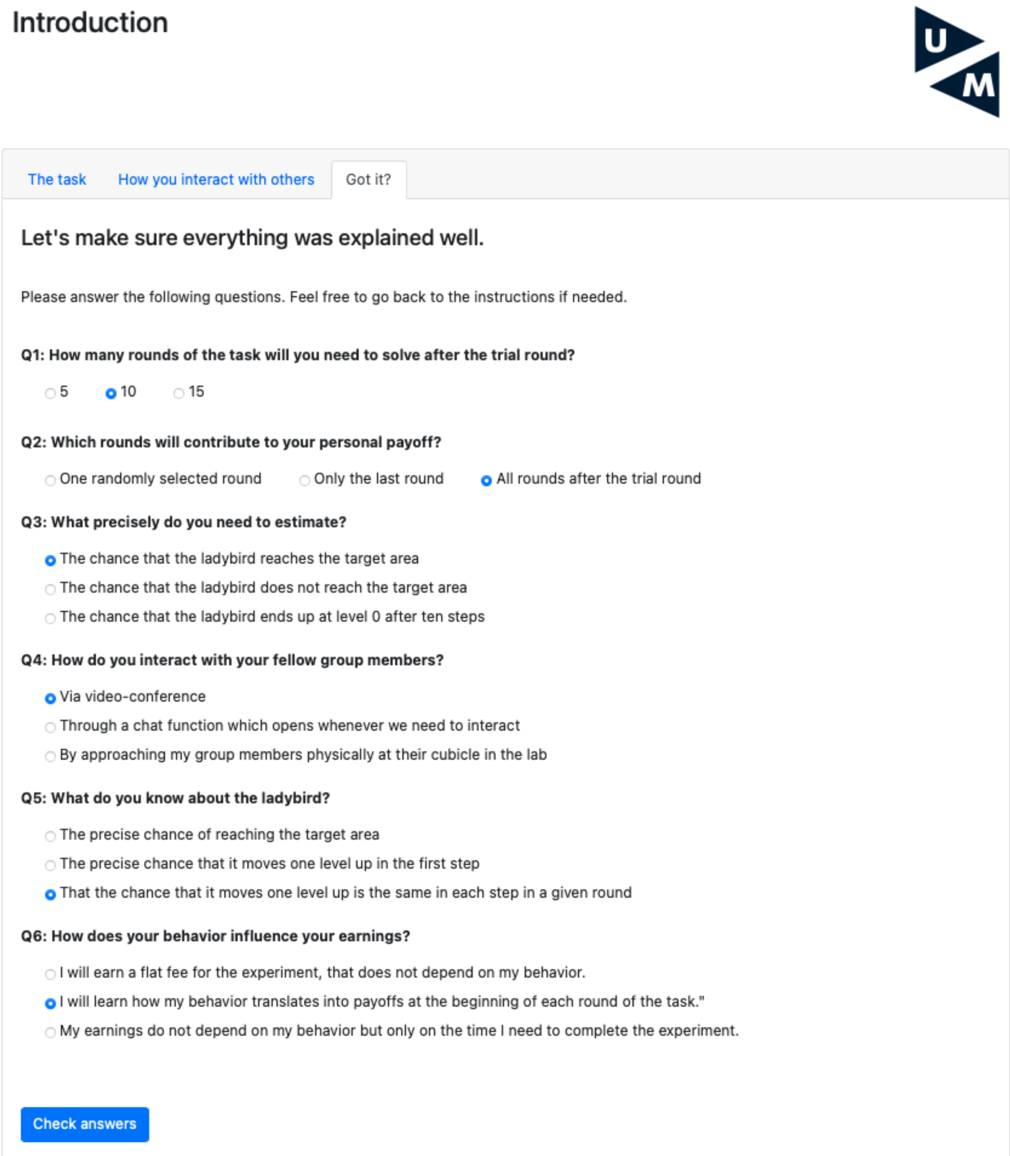
Notes: The introduction video on the judgment task can be found in the replication package.

Figure 16: 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 17: Check of understanding in FTF treatments



Introduction

The task | How you interact with others | Got it?

Let's make sure everything was explained well.

Please answer the following questions. Feel free to go back to the instructions if needed.

Q1: How many rounds of the task will you need to solve after the trial round?

- 5
- 10
- 15

Q2: Which rounds will contribute to your personal payoff?

- One randomly selected round
- Only the last round
- All rounds after the trial round

Q3: What precisely do you need to estimate?

- The chance that the ladybird reaches the target area
- The chance that the ladybird does not reach the target area
- The chance that the ladybird ends up at level 0 after ten steps

Q4: How do you interact with your fellow group members?

- Via video-conference
- Through a chat function which opens whenever we need to interact
- By approaching my group members physically at their cubicle in the lab

Q5: What do you know about the ladybird?

- The precise chance of reaching the target area
- The precise chance that it moves one level up in the first step
- That the chance that it moves one level up is the same in each step in a given round

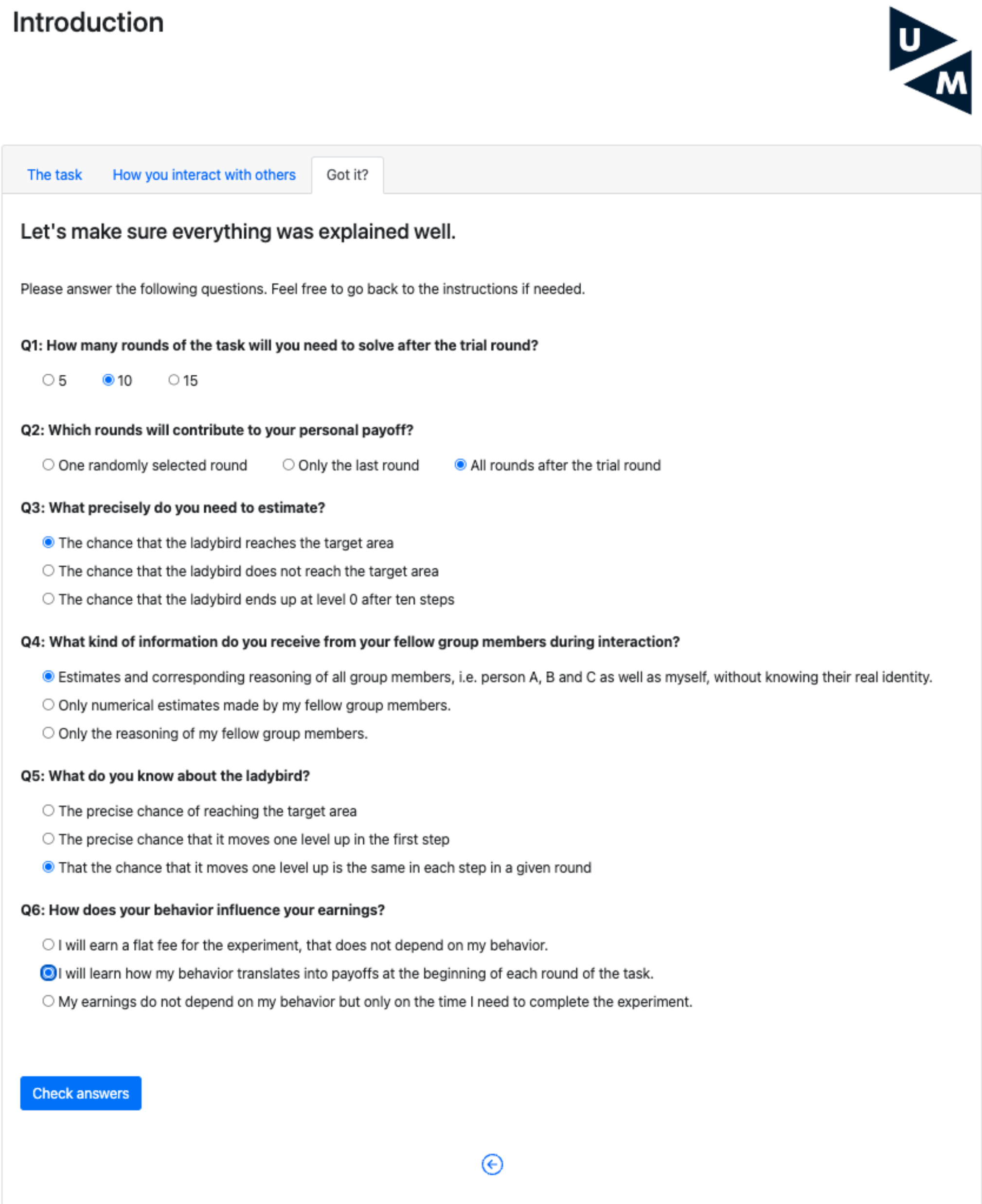
Q6: How does your behavior influence your earnings?

- I will earn a flat fee for the experiment, that does not depend on my behavior.
- I will learn how my behavior translates into payoffs at the beginning of each round of the task."
- My earnings do not depend on my behavior but only on the time I need to complete the experiment.

Check answers

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.

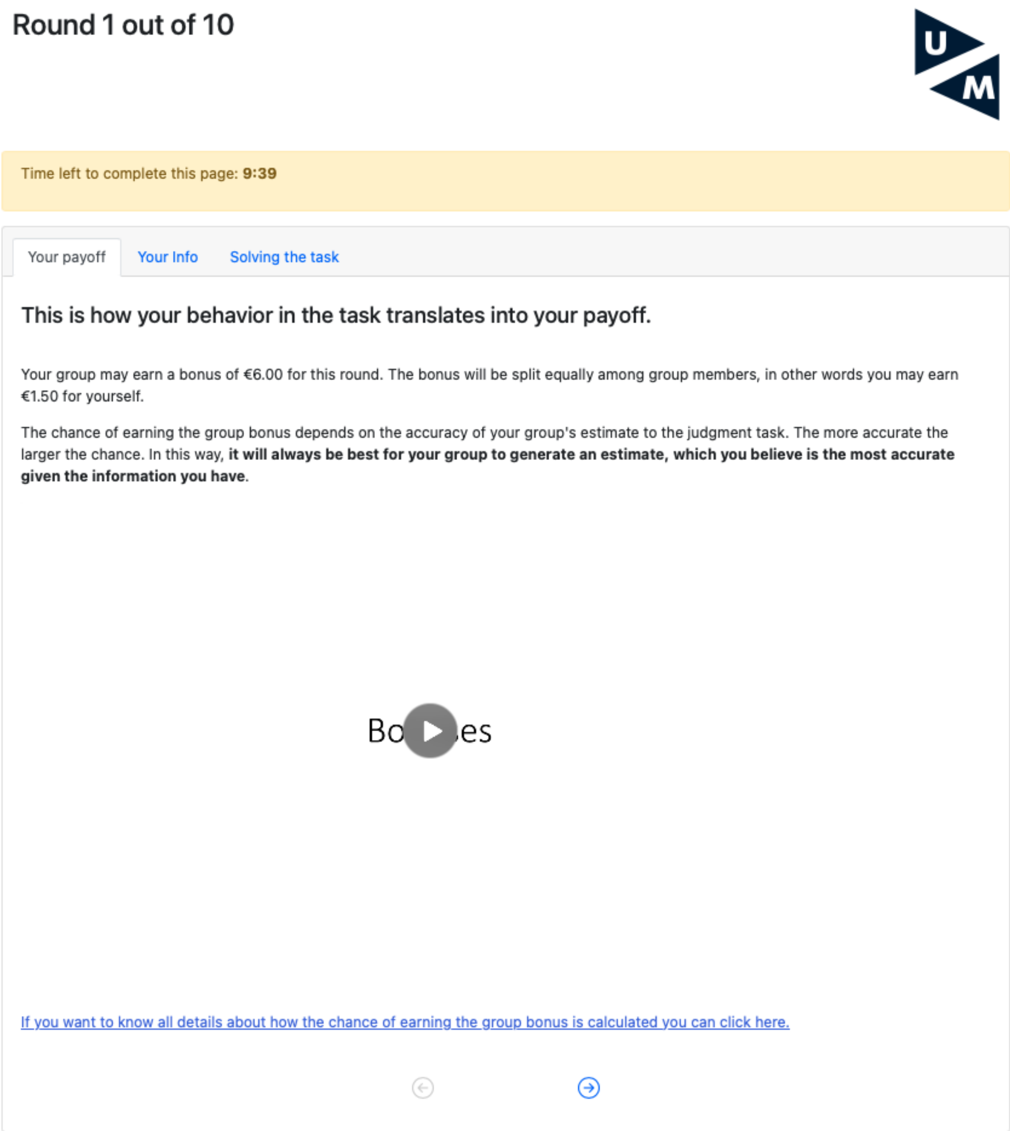
Figure 18: Check of understanding in Delphi treatments



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.

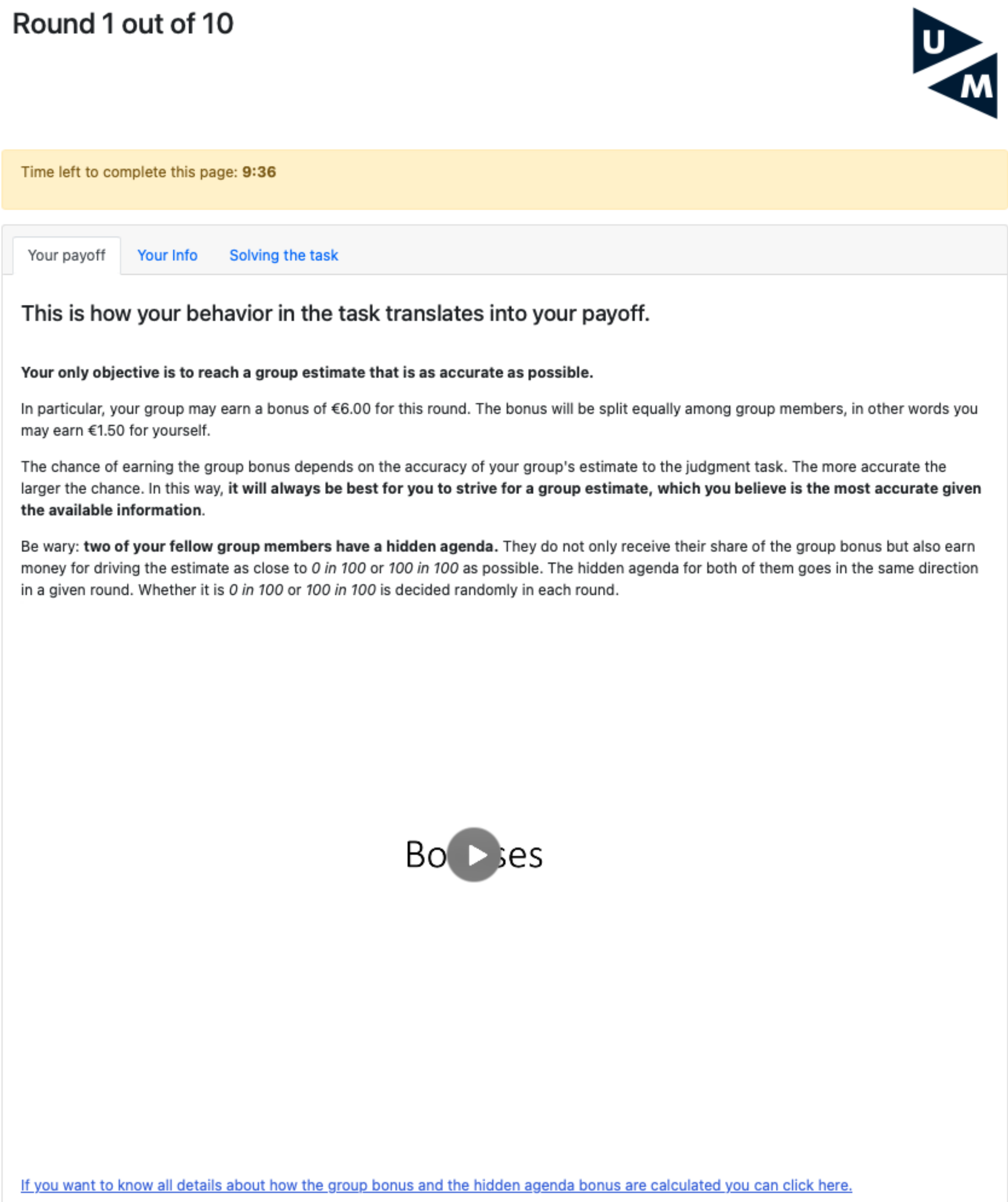
E.1.3 Judgment task

Figure 19: Introduction / repetition of payment information in FTF treatment without hidden agendas



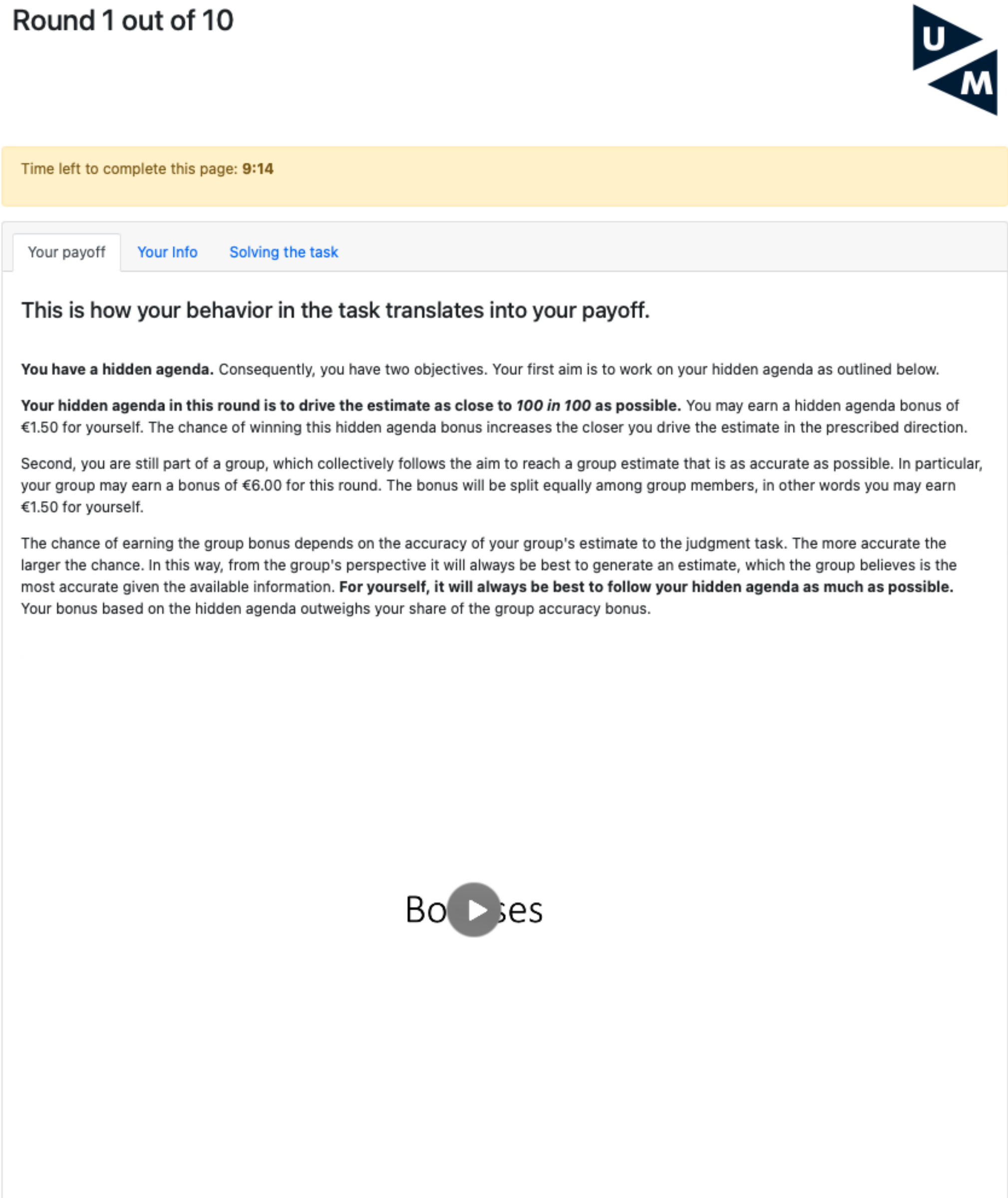
Notes: The corresponding introduction video on the bonus payments can be found in the replication package.

Figure 20: Introduction / repetition of payment information to participants without hidden agenda in FTF treatment with hidden agenda 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 with hidden agenda in FTF treatment with hidden agenda agendas



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 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.**

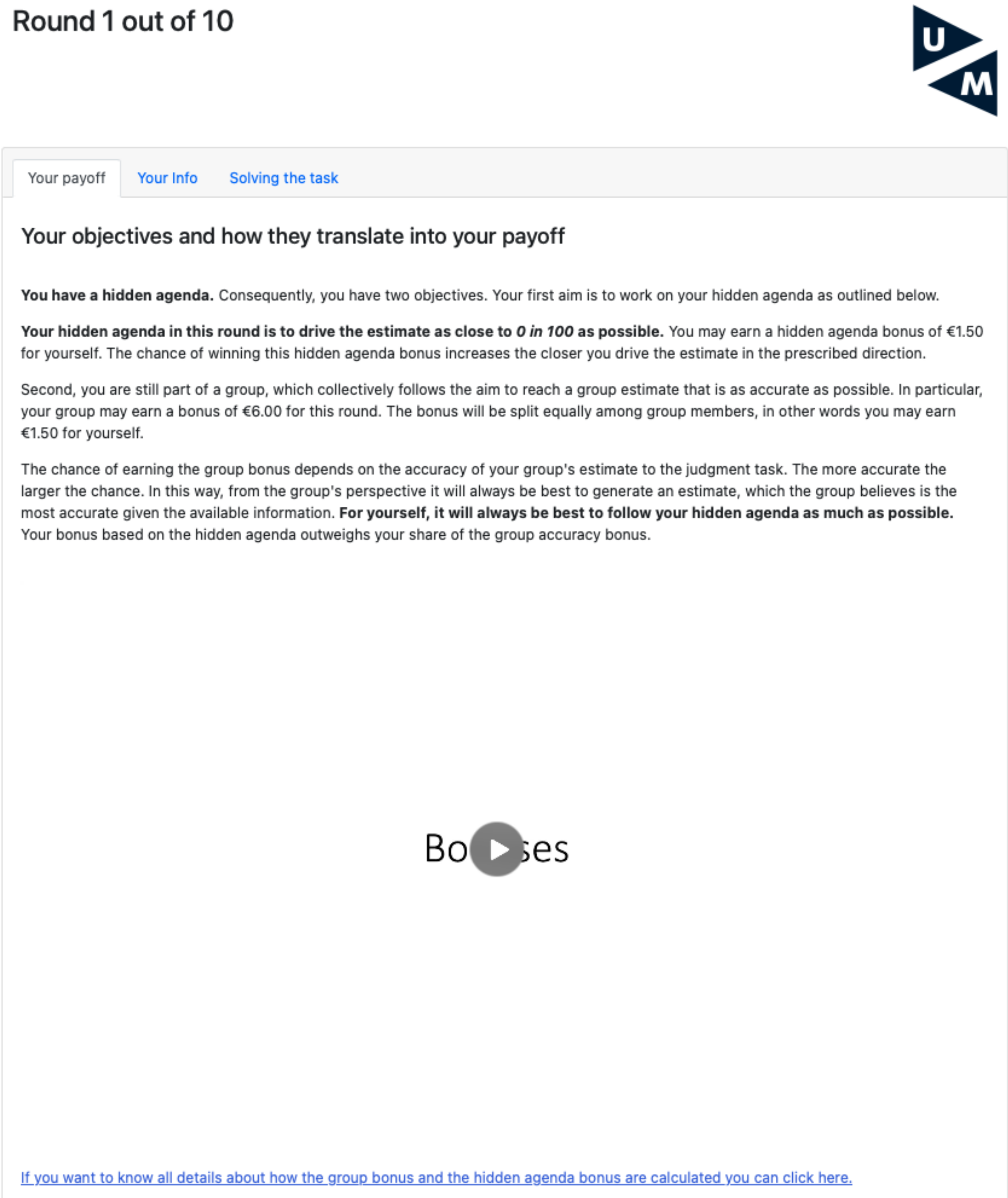
Be wary: **two of your fellow group members have a hidden agenda.** They do not only receive their share of the group bonus but also earn money for driving the estimate as close to *0 in 100* or *100 in 100* as possible. The hidden agenda for both of them goes in the same direction in a given round.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

Bo ses

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 23: Introduction / repetition of payment information to participants with 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

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.

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: Individual information on the judgment task in FTF treatments

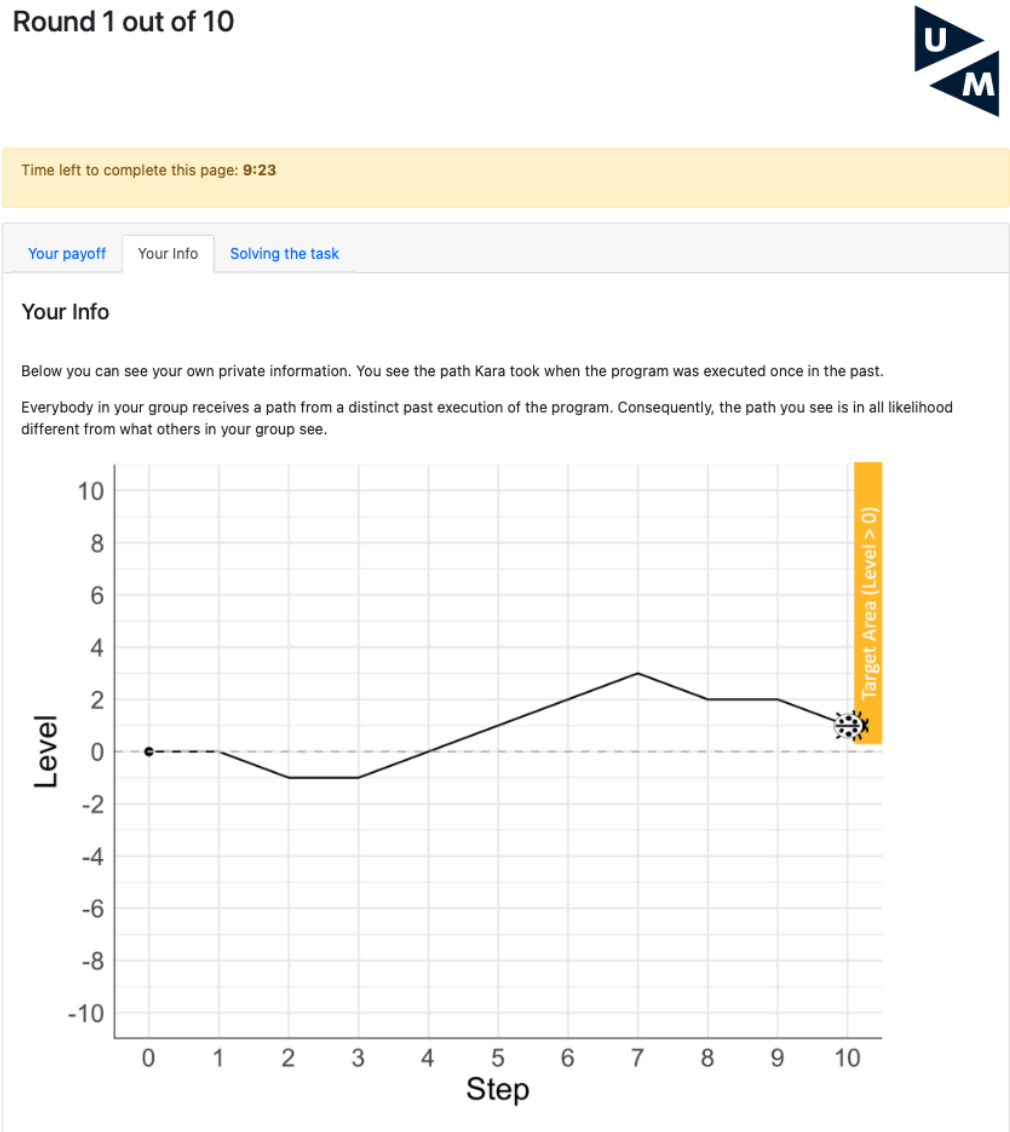


Figure 25: Individual information on the judgment task in Delphi treatments

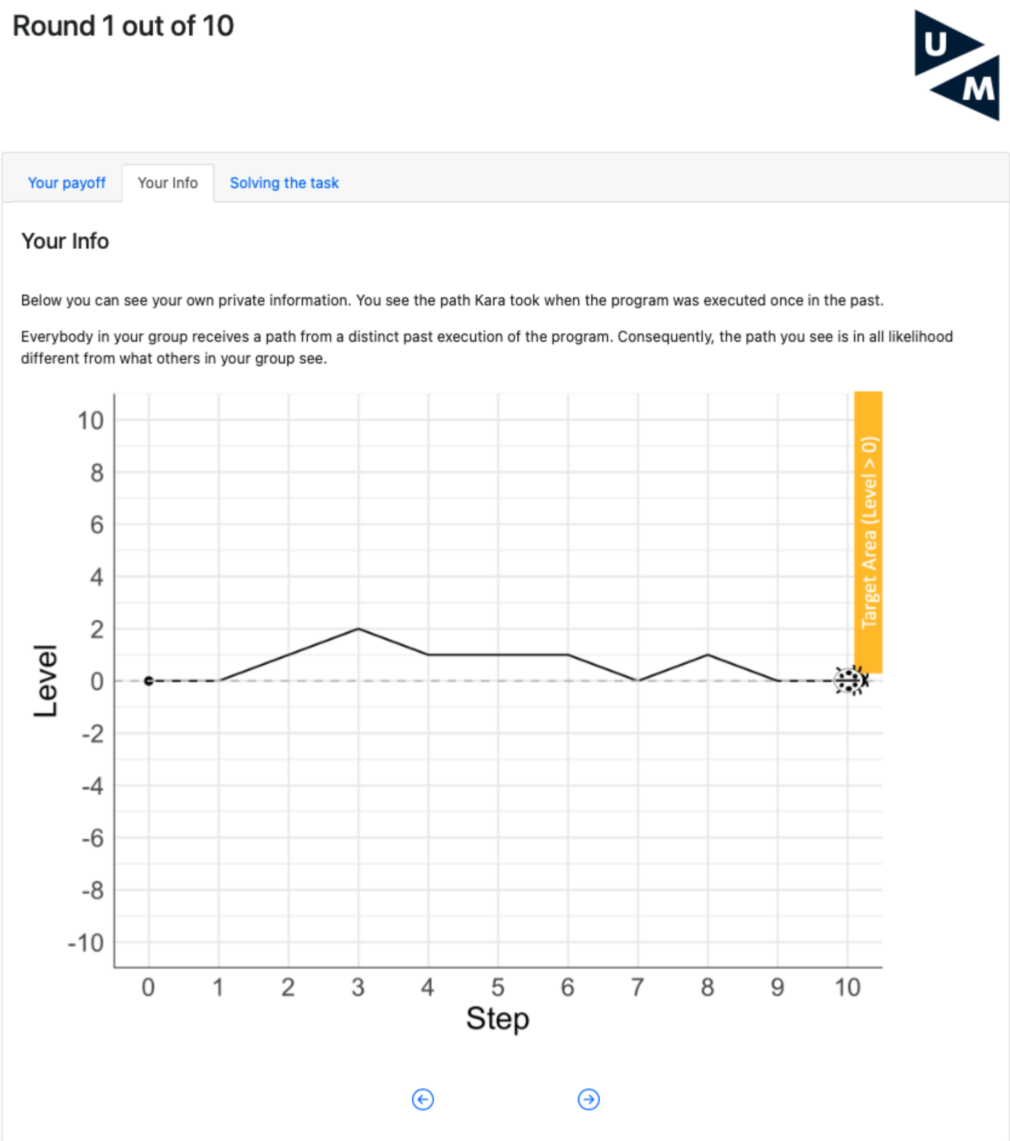


Figure 26: Input screen for group judgments in FTF treatments for participants without hidden agendas

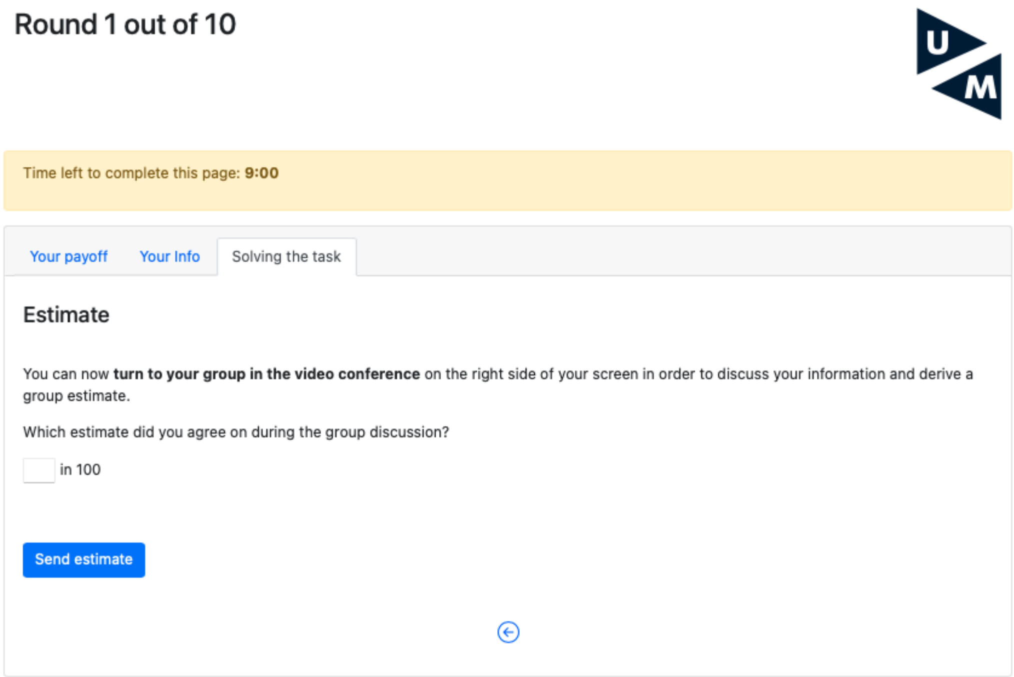


Figure 27: Input screen for group judgments for participants with hidden agenda in FTF treatment with hidden agendas

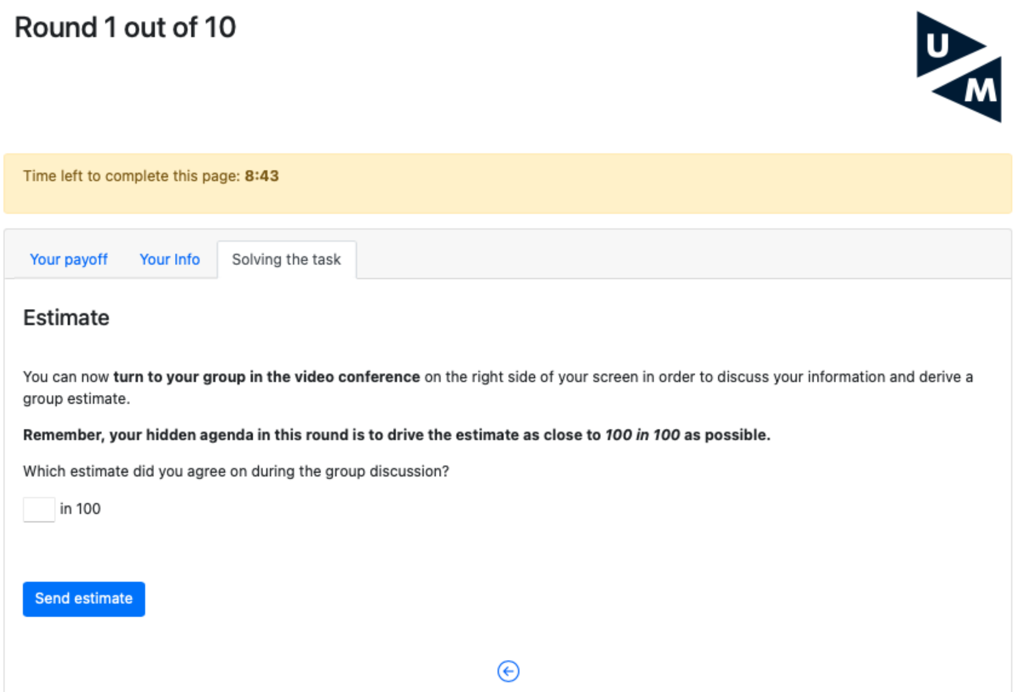


Figure 28: Input screen for first individual judgments for participants without hidden agenda in Delphi treatments

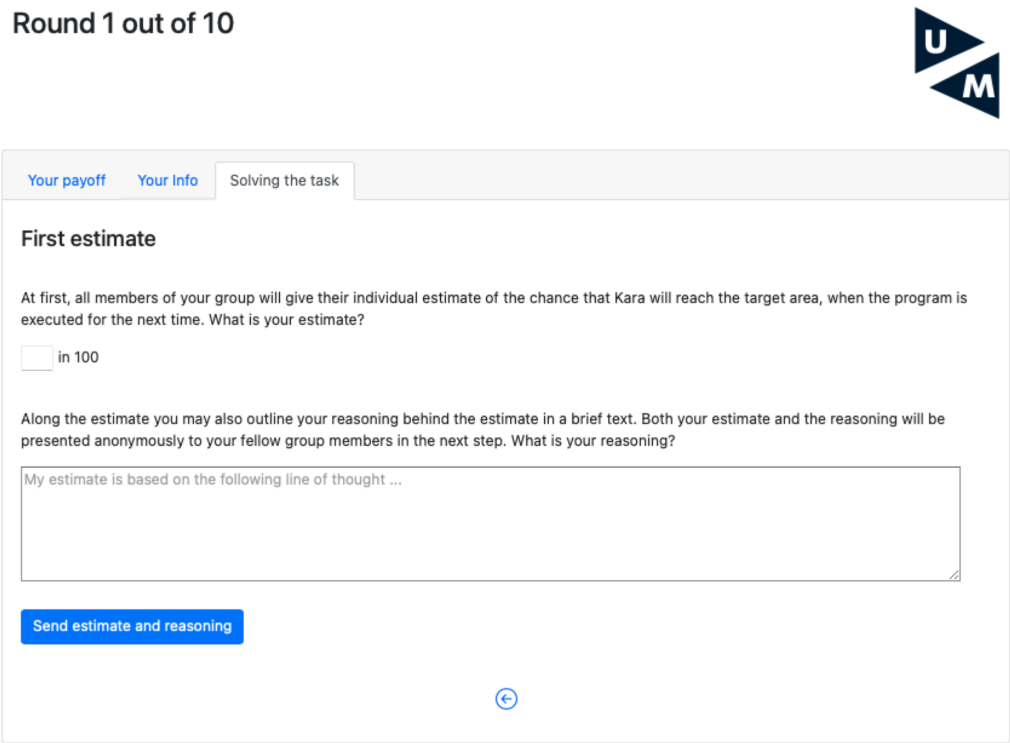


Figure 29: Input screen for first individual judgments for participants with hidden agenda in Delphi treatments with hidden agendas

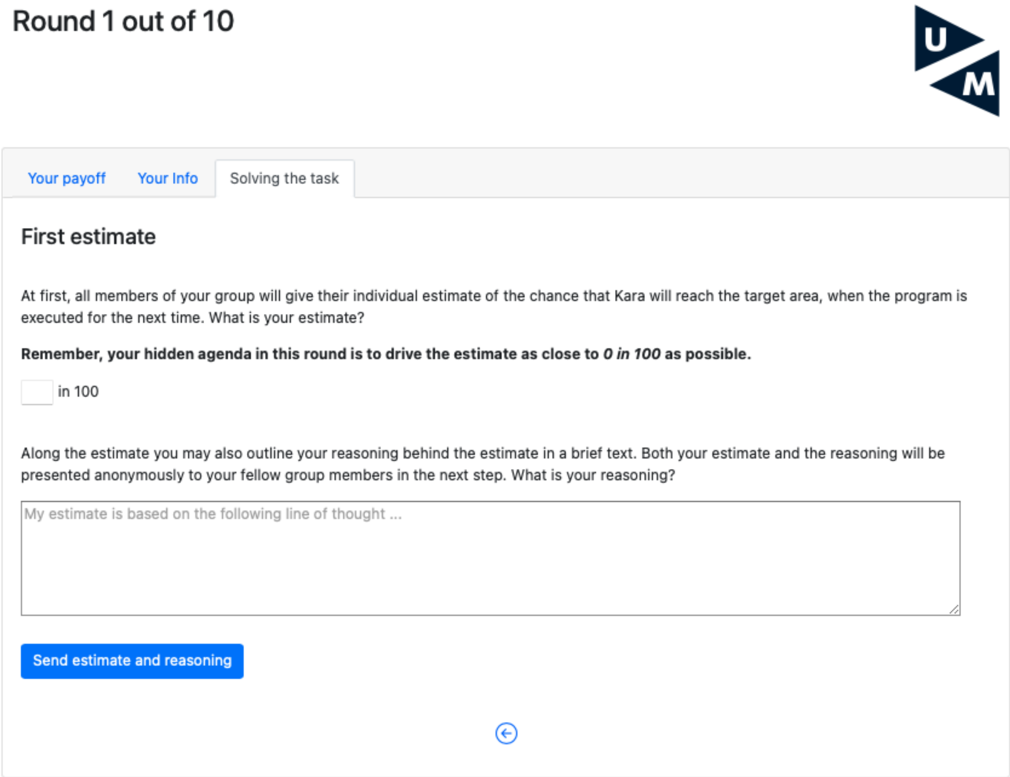


Figure 30: Feedback screen after first individual judgments for participants without hidden agenda in Delphi treatments

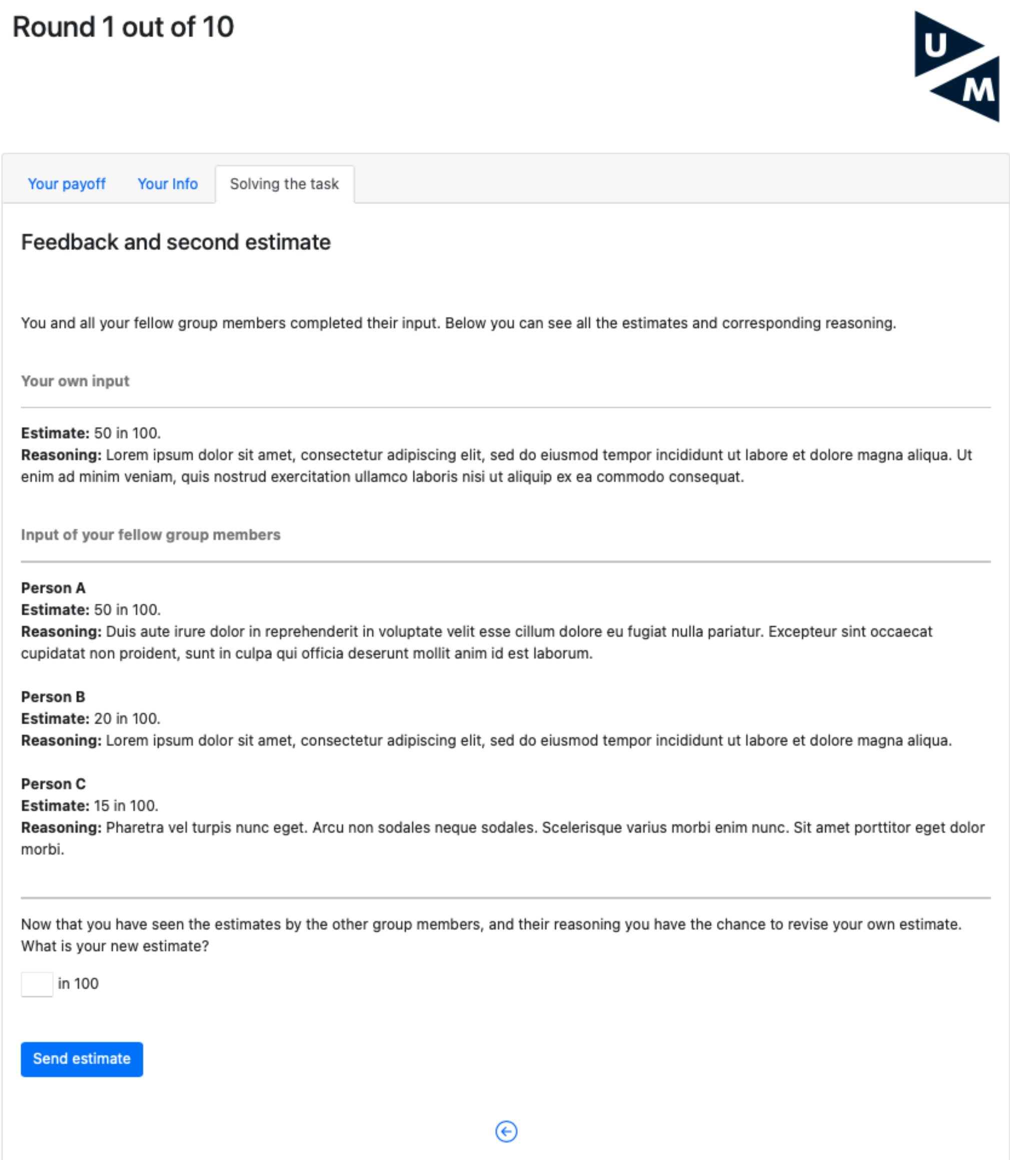
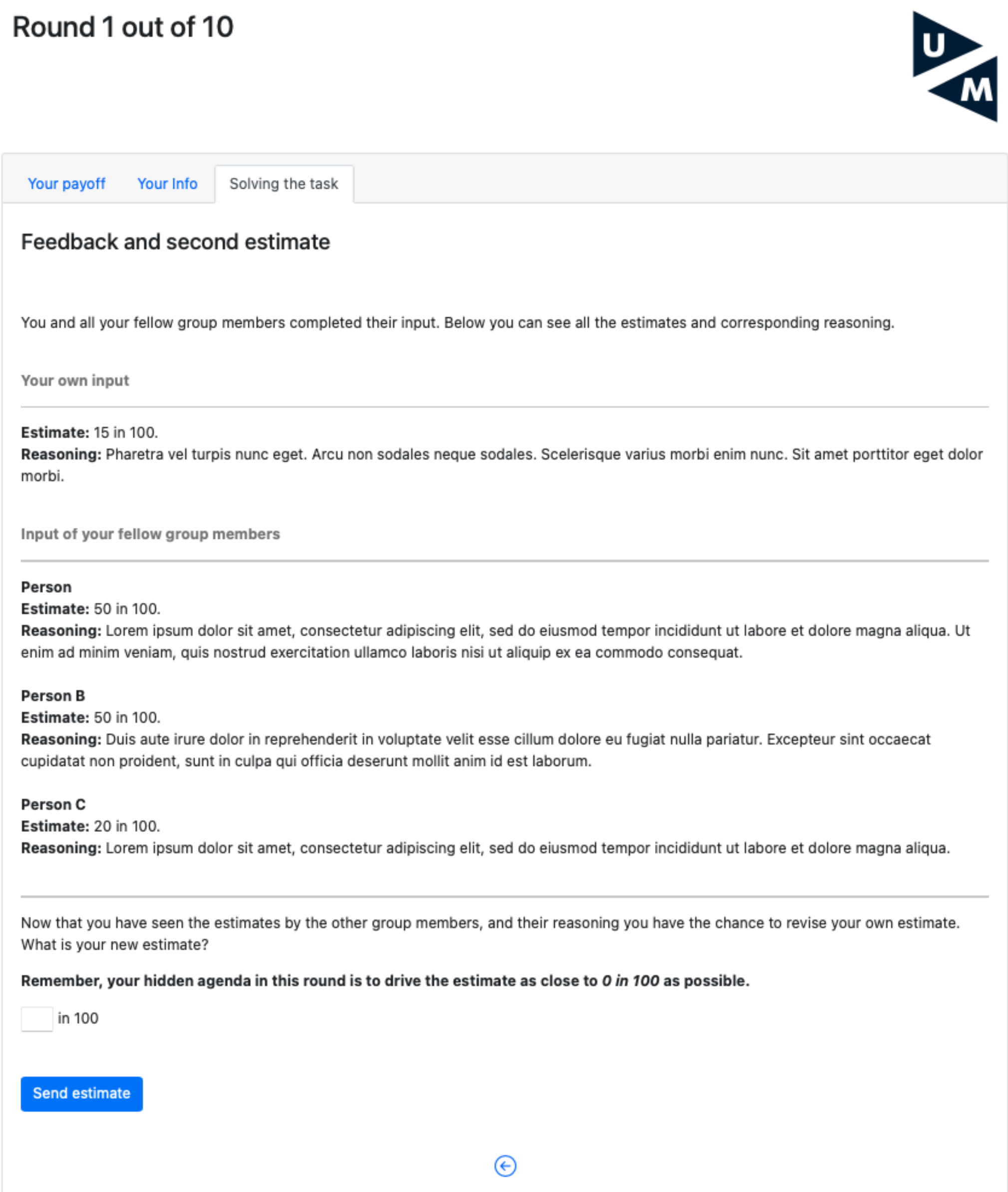


Figure 31: Feedback screen after first individual judgments for participants with hidden agenda in Delphi treatments with hidden agendas



Round 1 out of 10

Your payoff | Your Info | Solving the task

Feedback and second estimate

You and all your fellow group members completed their input. Below you can see all the estimates and corresponding reasoning.

Your own input

Estimate: 15 in 100.
Reasoning: Pharetra vel turpis nunc eget. Arcu non sodales neque sodales. Scelerisque varius morbi enim nunc. Sit amet porttitor eget dolor morbi.

Input of your fellow group members

Person
Estimate: 50 in 100.
Reasoning: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Person B
Estimate: 50 in 100.
Reasoning: Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Person C
Estimate: 20 in 100.
Reasoning: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

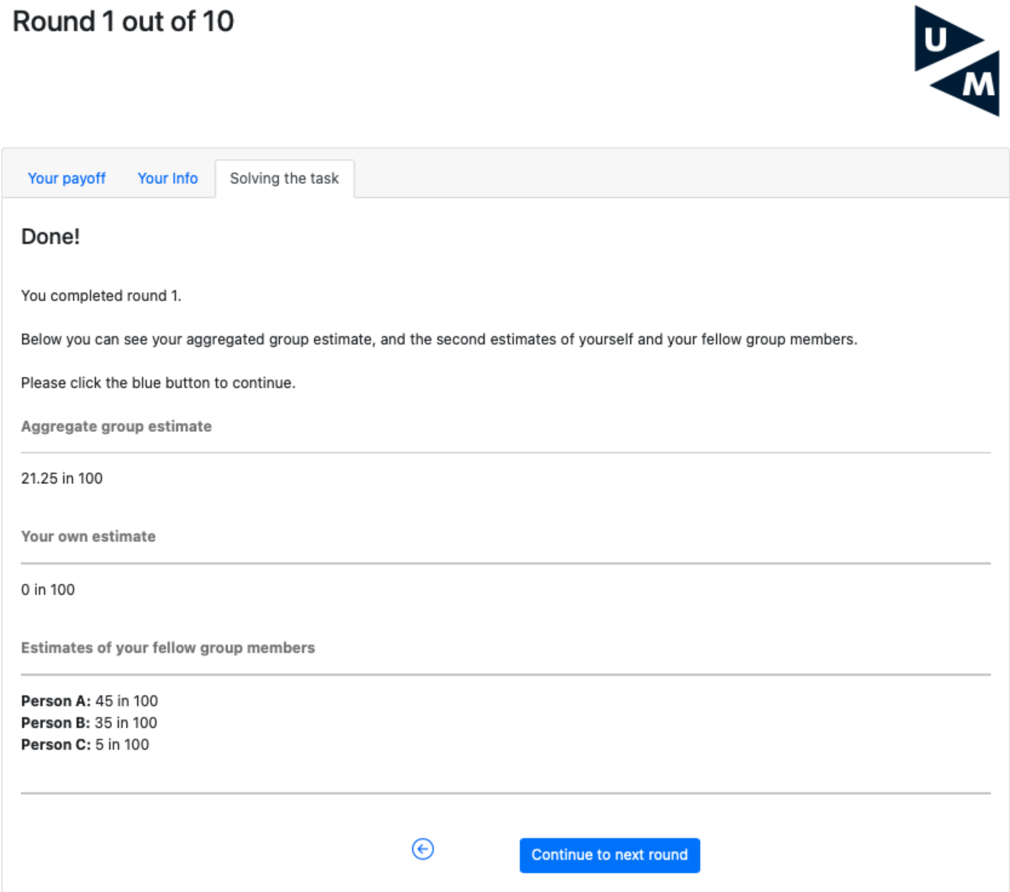
Now that you have seen the estimates by the other group members, and their reasoning you have the chance to revise your own estimate. What is your new estimate?

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

in 100

Send estimate

Figure 32: Feedback screen after second individual judgments for participants in Delphi treatments



E.1.4 Closing survey

Figure 33: Closing survey - personal characteristics

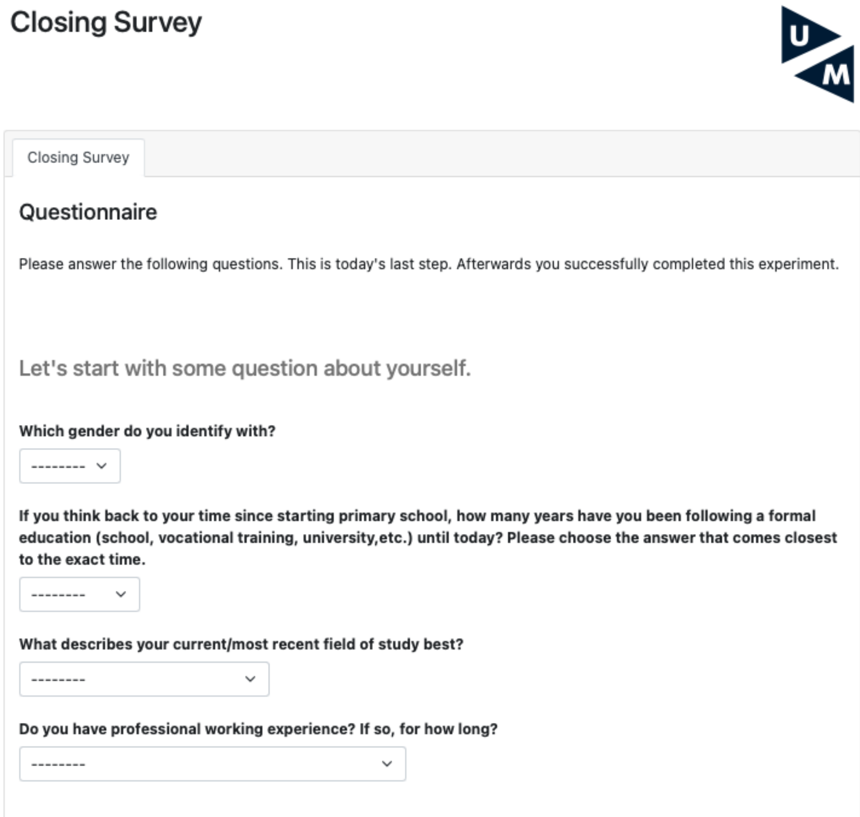
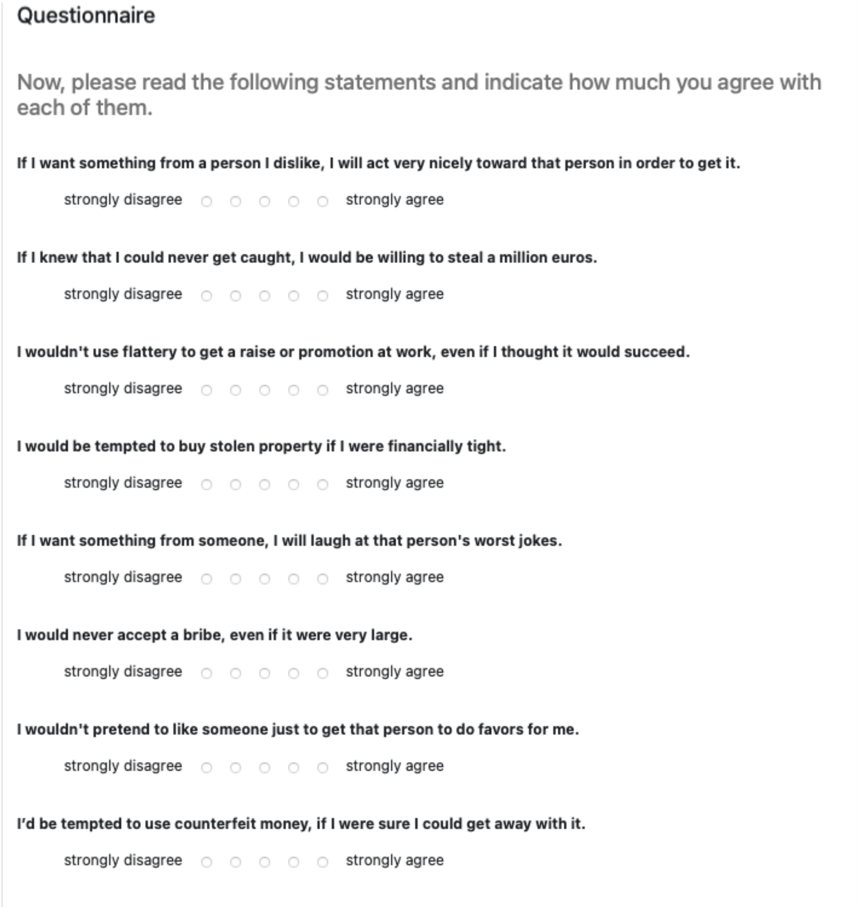


Figure 34: Closing survey - honesty module



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

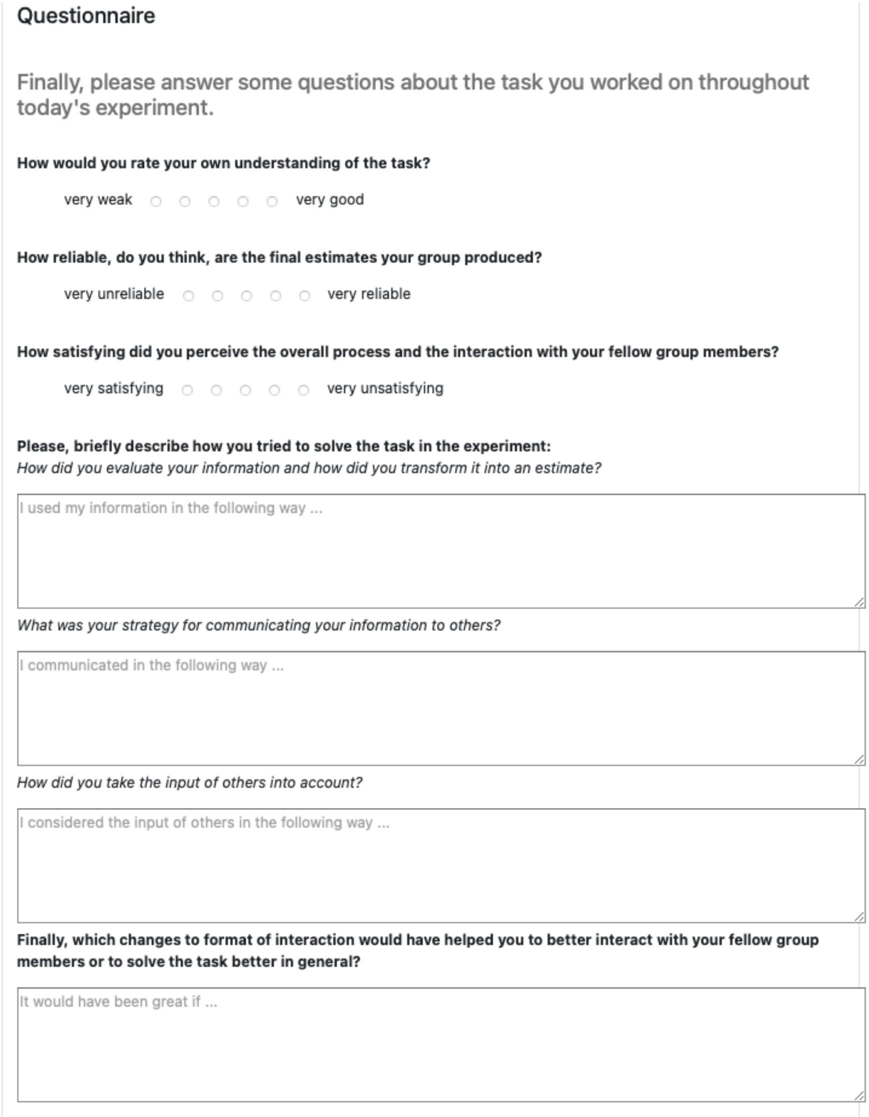
I wouldn't pretend to like someone just to get that person to do favors for me.

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 35: Closing survey - experience during experiment



Questionnaire

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

How reliable, do you think, are the final estimates your group produced?

very unreliable ○ ○ ○ ○ ○ very reliable

How satisfying did you perceive the overall process and the interaction with your fellow group members?

very satisfying ○ ○ ○ ○ ○ very unsatisfying

Please, briefly describe how you tried to solve the task in the experiment:
How did you evaluate your information and how did you transform it into an estimate?

I used my information in the following way ...

What was your strategy for communicating your information to others?

I communicated in the following way ...

How did you take the input of others into account?

I considered the input of others in the following way ...

Finally, which changes to format of interaction would have helped you to better interact with your fellow group members or to solve the task better in general?

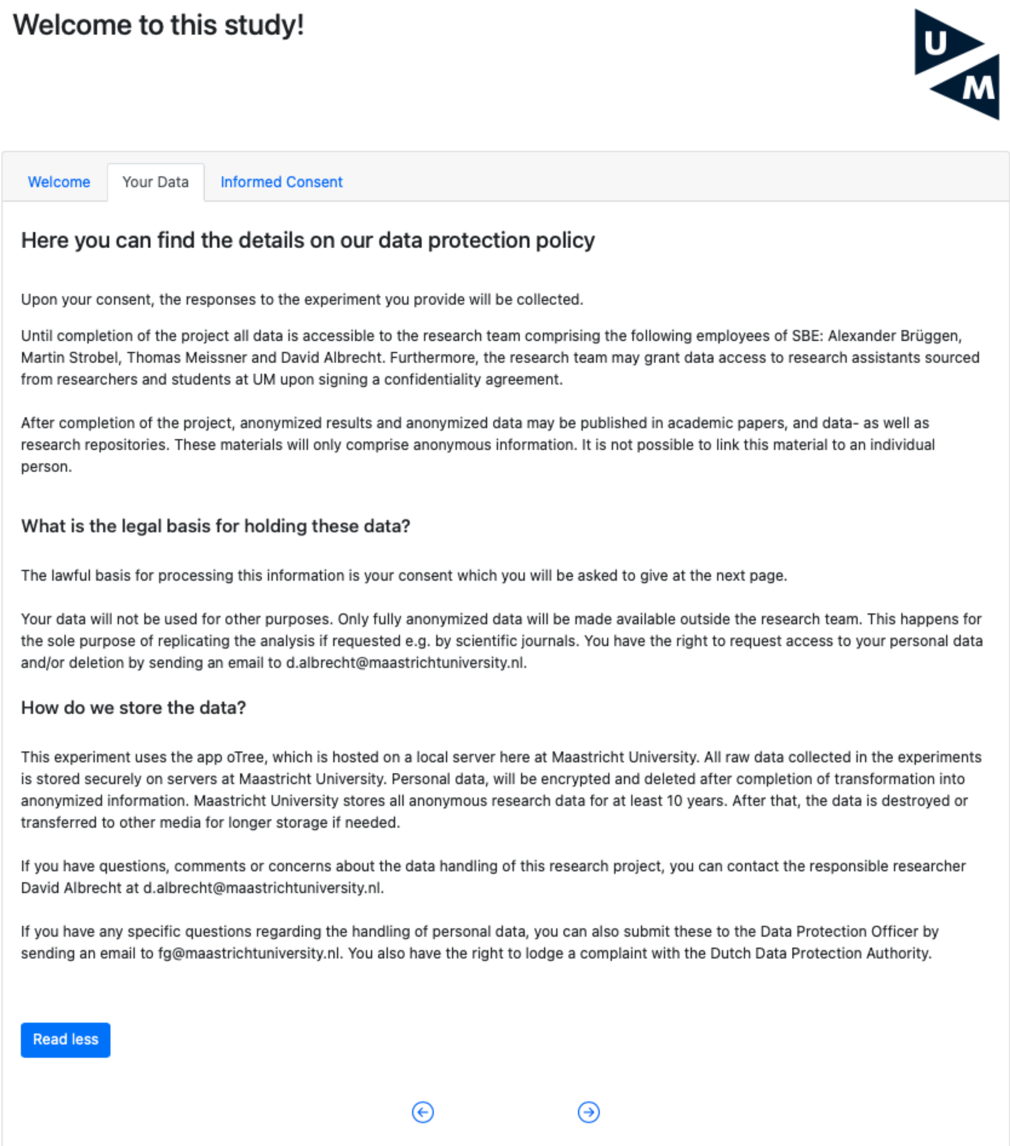
It would have been great if ...

E.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.

E.2.1 Welcome information

Figure 36: Opening screen of the decision experiment



Welcome to this study!

Welcome | Your Data | Informed Consent

Here you can find the details on our data protection policy

Upon your consent, the responses to the experiment you provide will be collected.

Until completion of the project all data is accessible to the research team comprising the following employees of SBE: Alexander Brüggen, Martin Strobel, Thomas Meissner and David Albrecht. Furthermore, the research team may grant data access to research assistants sourced from researchers and students at UM upon signing a confidentiality agreement.

After completion of the project, anonymized results and anonymized data may be published in academic papers, and data- as well as research repositories. These materials will only comprise anonymous information. It is not possible to link this material to an individual person.

What is the legal basis for holding these data?

The lawful basis for processing this information is your consent which you will be asked to give at the next page.

Your data will not be used for other purposes. Only fully anonymized data will be made available outside the research team. This happens for the sole purpose of replicating the analysis if requested e.g. by scientific journals. You have the right to request access to your personal data and/or deletion by sending an email to d.albrecht@maastrichtuniversity.nl.

How do we store the data?

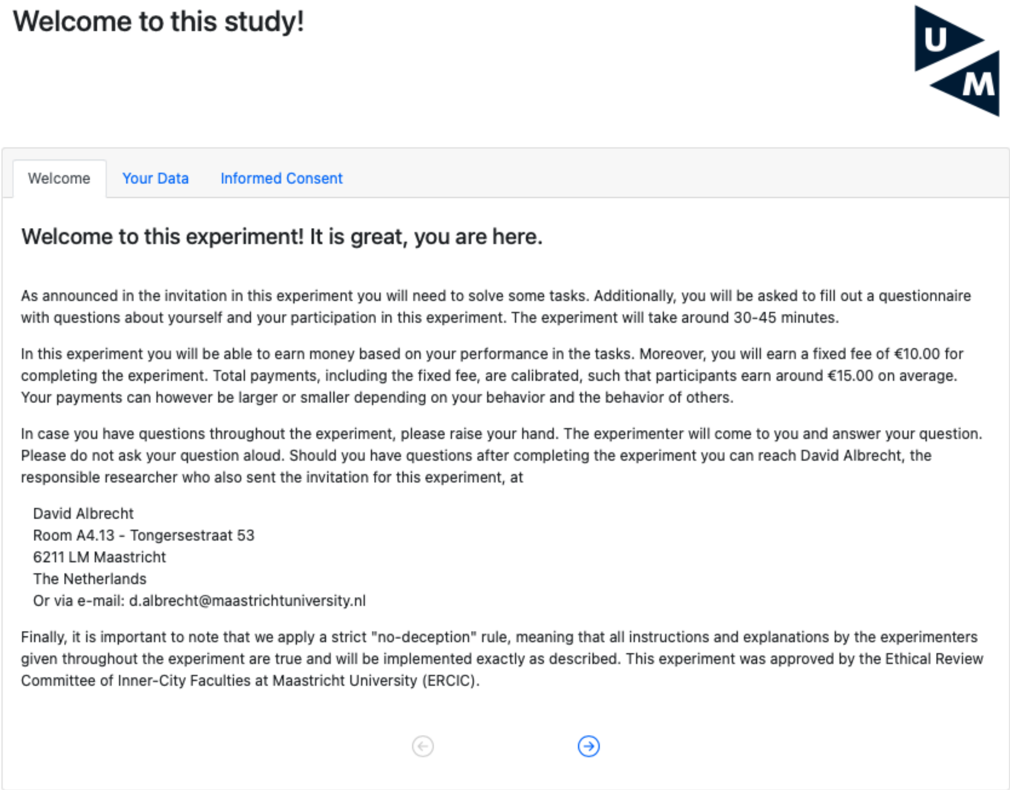
This experiment uses the app oTree, which is hosted on a local server here at Maastricht University. All raw data collected in the experiments is stored securely on servers at Maastricht University. Personal data, will be encrypted and deleted after completion of transformation into anonymized information. Maastricht University stores all anonymous research data for at least 10 years. After that, the data is destroyed or transferred to other media for longer storage if needed.

If you have questions, comments or concerns about the data handling of this research project, you can contact the responsible researcher David Albrecht at d.albrecht@maastrichtuniversity.nl.

If you have any specific questions regarding the handling of personal data, you can also submit these to the Data Protection Officer by sending an email to fg@maastrichtuniversity.nl. You also have the right to lodge a complaint with the Dutch Data Protection Authority.

Read less

Figure 37: 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.

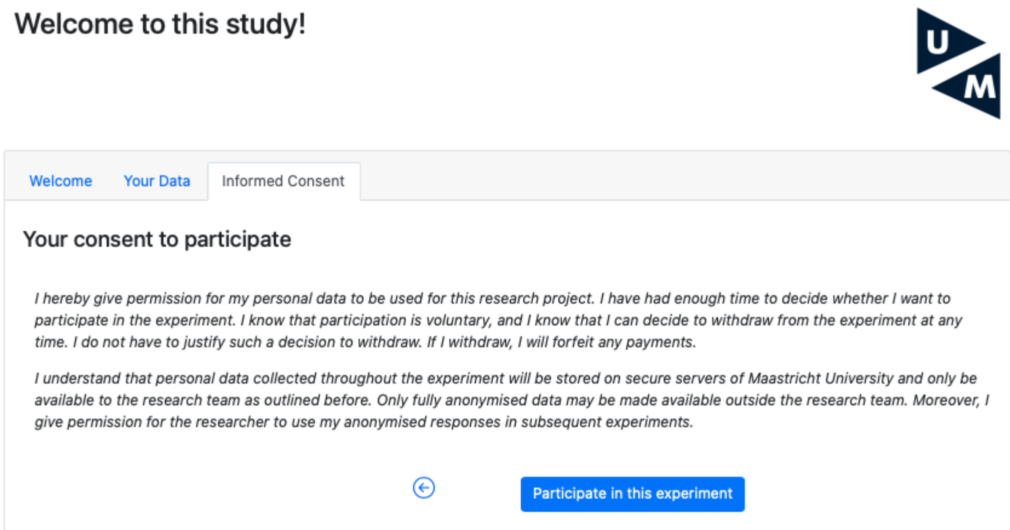
In this experiment you will be able to earn money based on your performance in the tasks. Moreover, you will earn a fixed fee of €10.00 for completing the experiment. Total payments, including the fixed fee, are calibrated, such that participants earn around €15.00 on average. Your payments can however 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 38: Informed consent in the estimation experiment



Welcome to this study!

Welcome | Your Data | Informed Consent

Your consent to participate

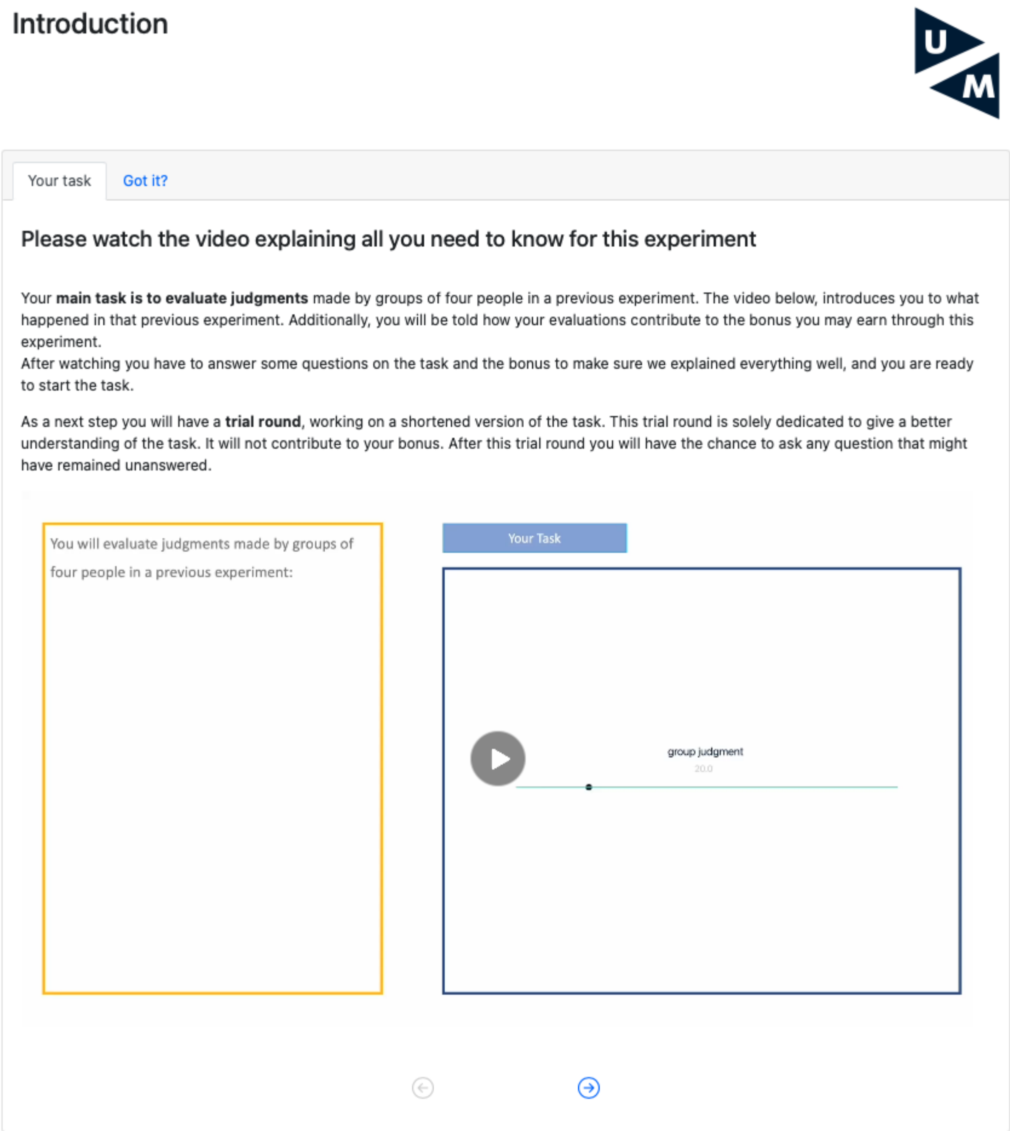
I hereby give permission for my personal data to be used for this research project. I have had enough time to decide whether I want to participate in the experiment. I know that participation is voluntary, and I know that I can decide to withdraw from the experiment at any time. I do not have to justify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

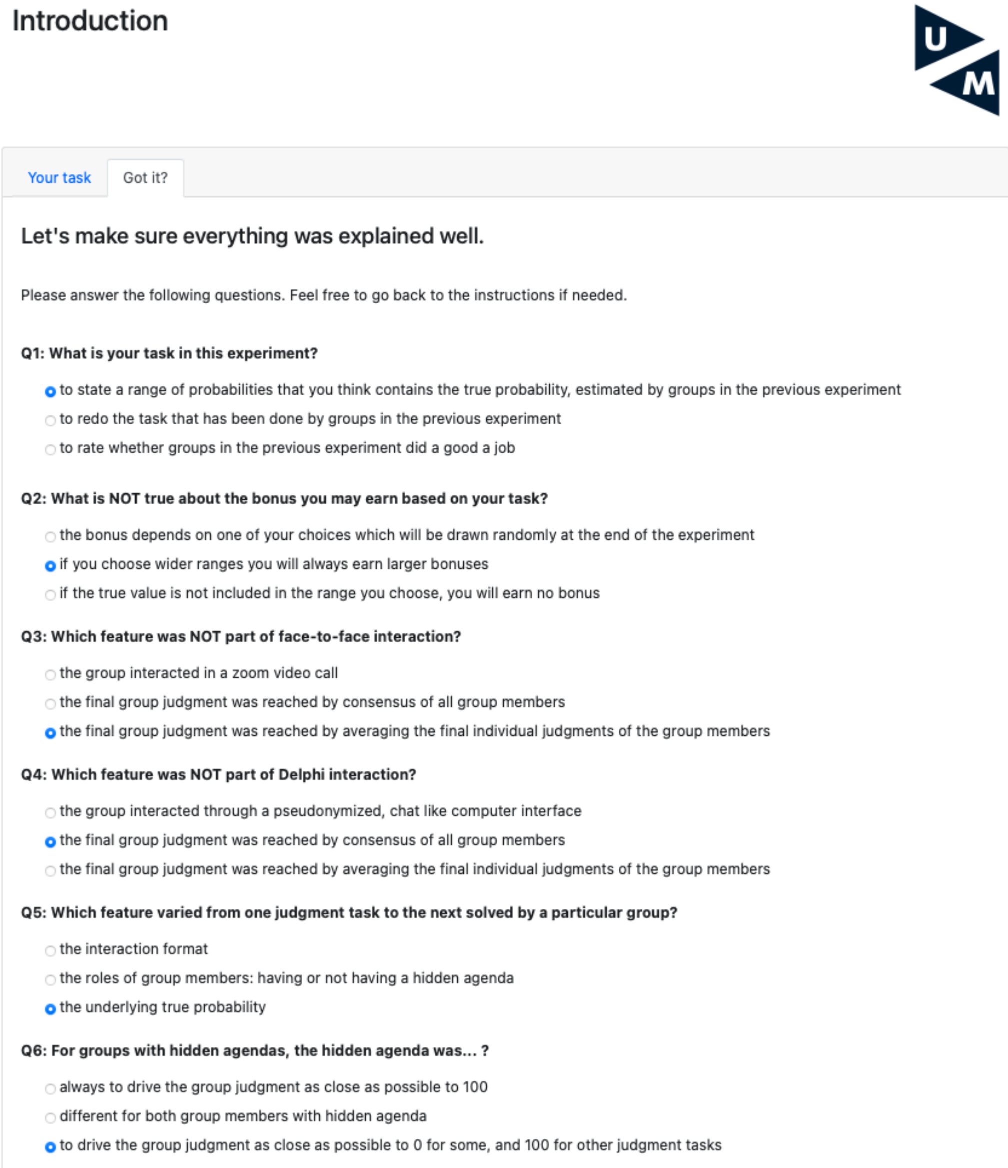
E.2.2 Introduction to the experiment

Figure 39: Introduction to the decision task



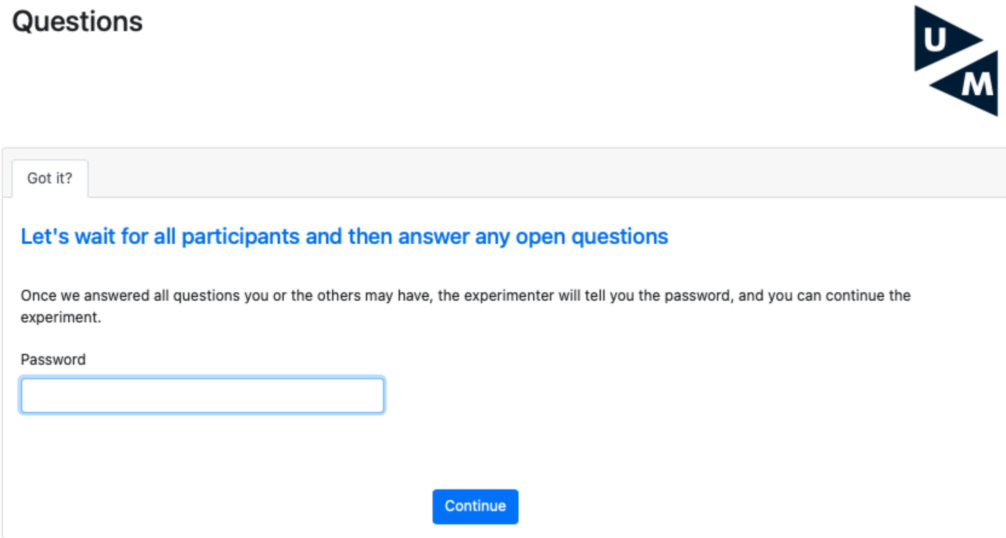
Notes: The introduction video on the decision task can be found in the replication package.

Figure 40: 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 41: Prompt to ask any question that may have remained unanswered after the introduction



E.2.3 Decision task

Figure 42: Decision task screen for evaluating group judgments from FTF groups without hidden agendas

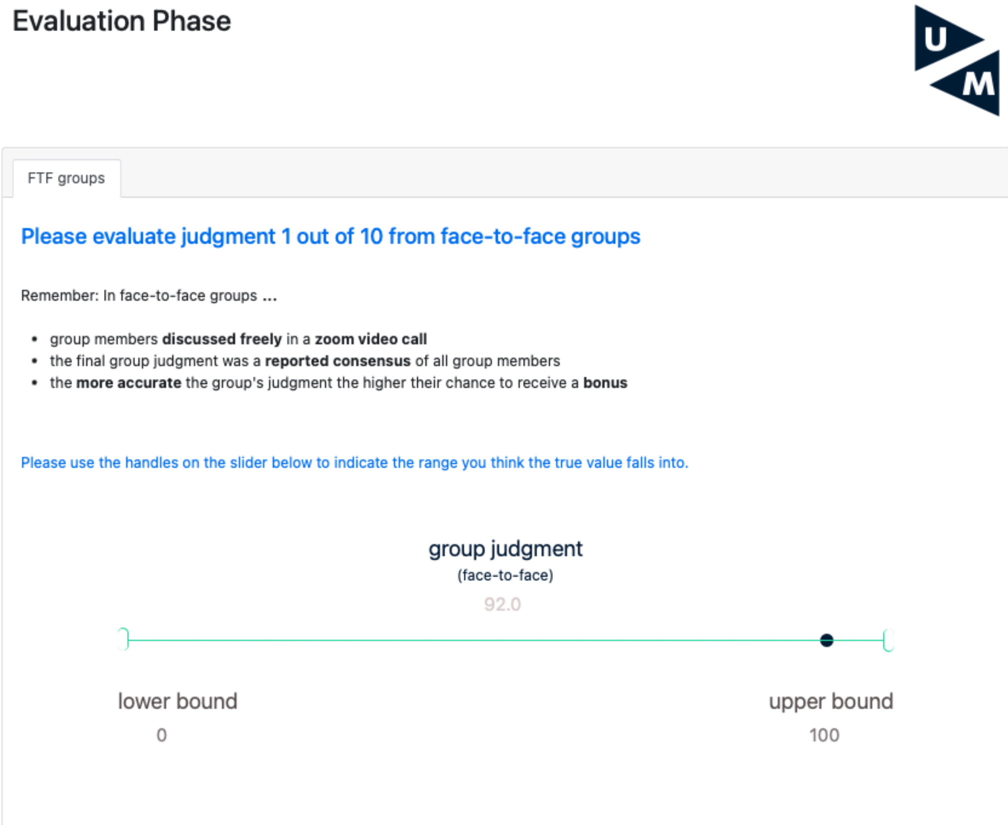


Figure 43: Decision task screen for evaluating group judgments from FTF groups with hidden agendas

Figure 44: Decision task screen for evaluating group judgments from Delphi groups without hidden agendas

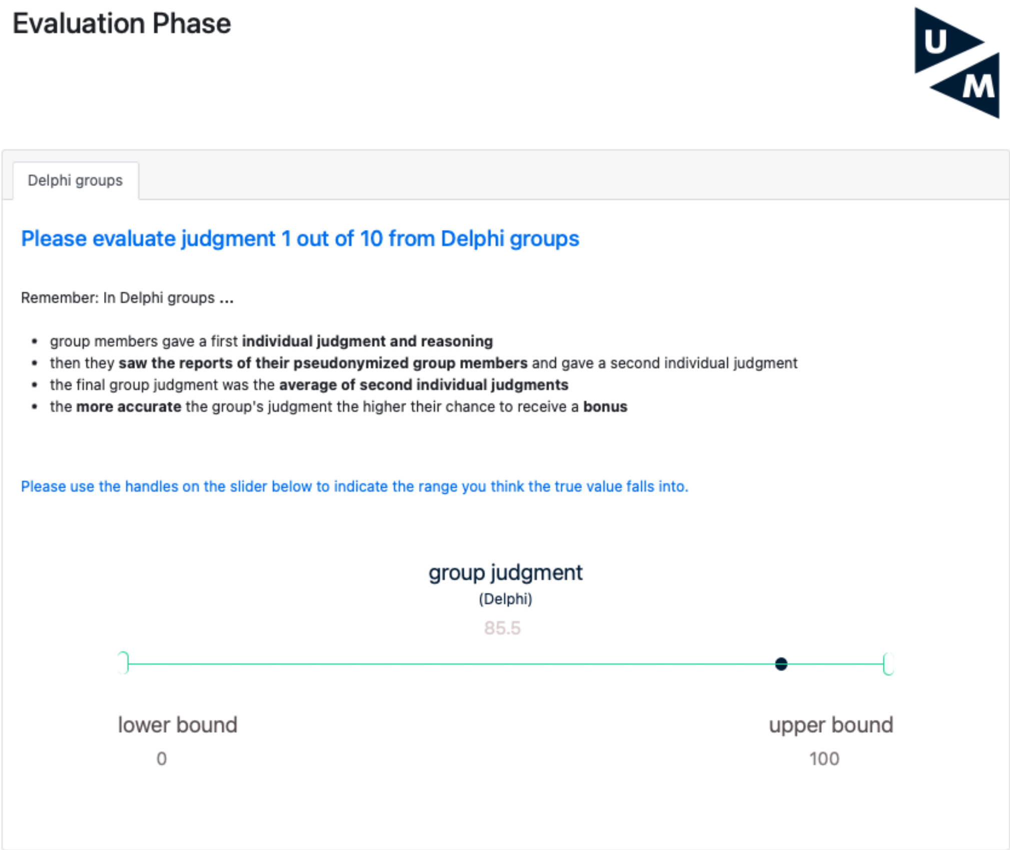
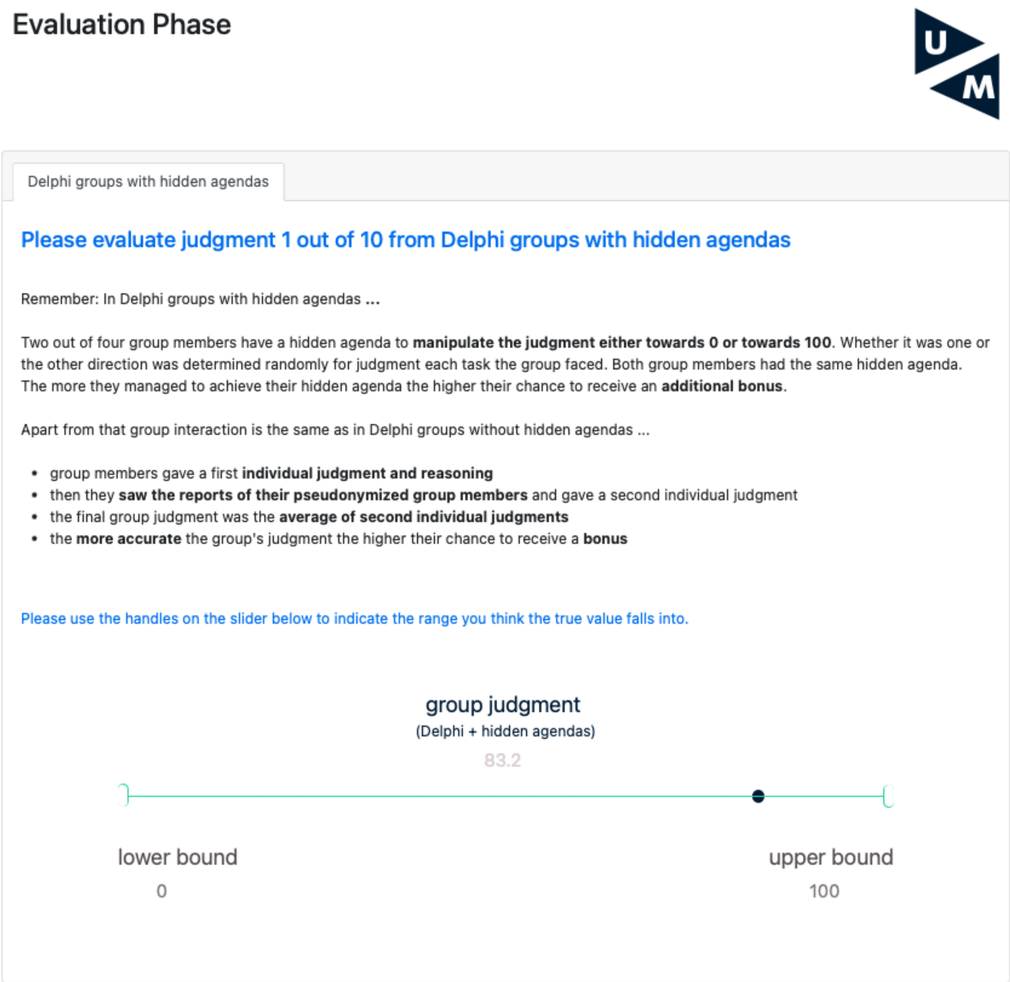
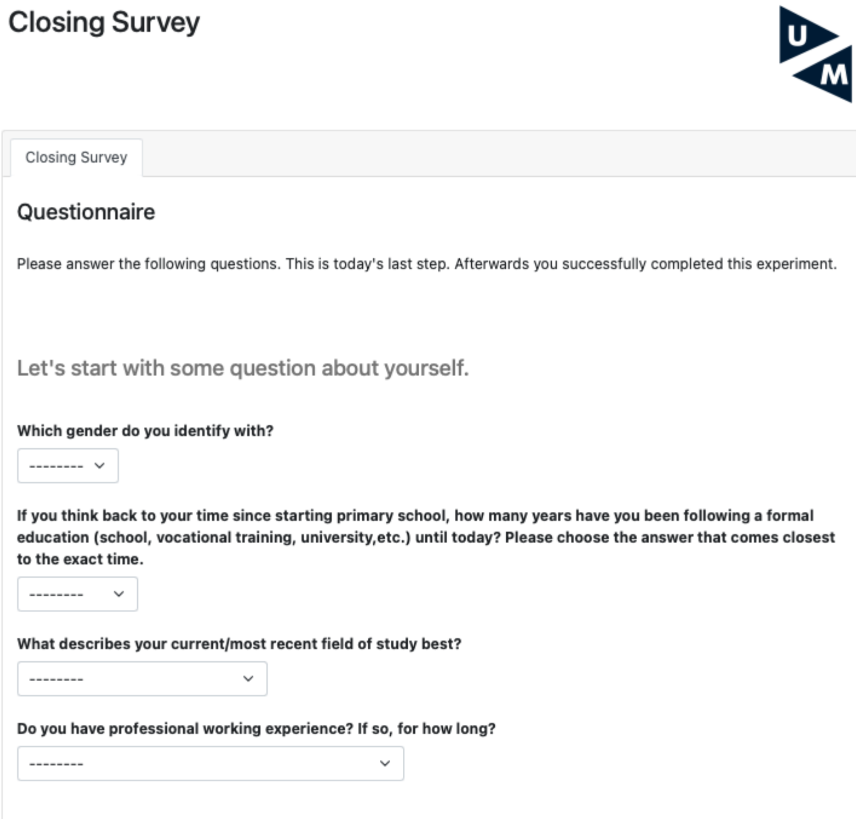


Figure 45: Decision task screen for evaluating group judgments from Delphi groups with hidden agendas



E.2.4 Closing survey

Figure 46: Closing survey - personal characteristics



Closing Survey

Closing Survey

Questionnaire

Please answer the following questions. This is today's last step. Afterwards you successfully completed this experiment.

Let's start with some question about yourself.

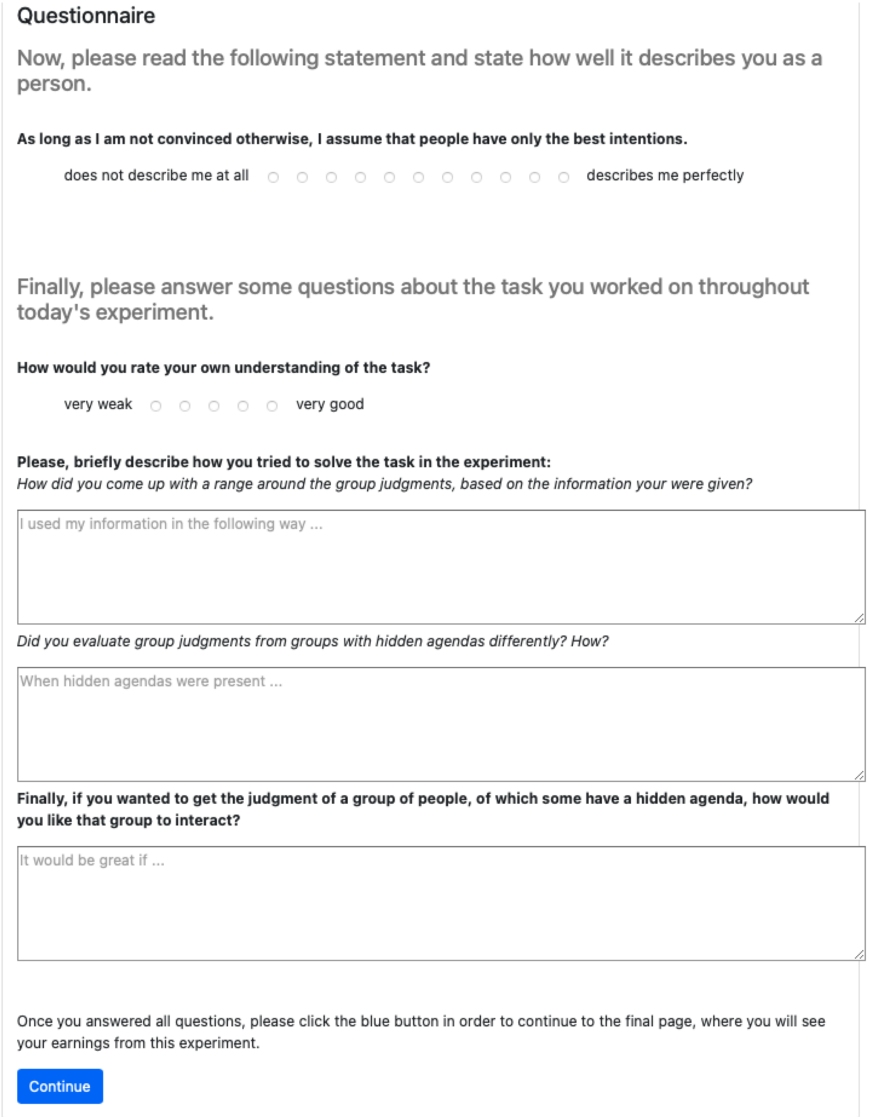
Which gender do you identify with?

If you think back to your time since starting primary school, how many years have you been following a formal education (school, vocational training, university,etc.) until today? Please choose the answer that comes closest to the exact time.

What describes your current/most recent field of study best?

Do you have professional working experience? If so, for how long?

Figure 47: 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