NET WORK LISER

Discussion Paper Series

IZA DP No. 19100

November 2026

Team Composition Matters for Overcoming Correlation Neglect in Belief Formation

Sofia Monteiro
Kiel Institute for the World Economy

Matthias Praxmarer
Max Planck Institute for Behavioral Economics

Stefan Schmidt
Max Planck Institute for Behavioral Economics

Matthias Sutter
Max Planck Institute for Behavioral Economics,
University of Cologne, University of Innsbruck
and IZA@LISER

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

Team Composition Matters for Overcoming Correlation Neglect in Belief Formation*

Abstract

Correlated signals are common in the information we process daily, and yet many people treat them as if they were independent, generating biased beliefs. Given that many important decisions are taken by teams, we study in an experiment with 410 participants whether team decision-making reduces this bias in comparison to individual decision making. At first sight, the answer is negative, at least in the aggregate, but closer inspection reveals that teams are less affected by correlation neglect and generate significantly higher earnings in the task. Most importantly, team composition matters crucially. Classifying team members as either naive (i.e., prone to correlation neglect) or rational, we show that having a rational partner halves a naive member's naivete, and the share of tasks in which a pair states the rational belief rises from 9% with two naive members to 40% with one rational member and 67% with two. Pairs with one rational member are, on average, about as naive as rational individuals, yet they fall short of a demanding "truth-wins norm", according to which teams should make rational choices whenever at least one member is classified as rational. Our analysis of team chats helps explain these patterns.

JEL classification

C91, D91, D83, D84

Keywords

beliefs, correlation neglect, bounded rationality, team decision-making, experiment

Corresponding author

Matthias Sutter
matthias.sutter@econ.mpg.de

* We would like to thank Michael Seebauer, Vyoma Muralidhara, Aurelia Ziegler, Cecilia Zölch, Michael Kuelper, Celina Kullmann and Alina Sowa for help in running the experiments. Bettina Rockenbach and Daniel Wiesen provided helpful comments. Financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC 2126/2–390838866 — is gratefully acknowledged. The study has IRB-approval (2018-3 and 2021-36) from the Max Planck Institute for Behavioral Economics.

1 Introduction

Human decision making depends upon information about the state of the world. For this reason, collecting information and generating beliefs about the world on the basis of this information is important. Yet, different pieces of information need not necessarily be independent of each other, but rather they may be correlated (Enke and Zimmermann, 2019). If this is the case, human decision making can be prone to bias by lending too much weight on correlated information, without backing out the underlying common source of it. These biases can be costly to both the individual and our society when making choices for the economy, in an election, or in health-related issues, if the set of information considered treats correlated information as independent.

One illustrative example for correlated information is the observation that CNN, Fox News and Sky News all share a common source for their information, i.e., Reuters. Watching these TV channels, while at the same time disregarding their common information source, may therefore induce the audience to belief that different sources confirmed the same content, thus making it presumably more credible. Correlated signals may thus have a larger weight on the formation of beliefs than warranted. The same applies in social networks where individual opinions of different network members or friends are often partially based on information from shared acquaintances. This means that when network members interact with each other, they are also acting on correlated information where, for example, the same stories are being retold multiple times (Akerlof and Shiller, 2009).

Enke and Zimmermann (2019) have shown that many subjects suffer in their belief formation from correlation neglect, i.e., they are unable to correct for the correlational nature of different pieces of information when forming beliefs about the true state of the world in their experiment. In their seminal study, they present evidence that naive beliefs are not driven by inadequate mathematical and computational skills necessary to process correlated information. Rather, subjects show conceptual problems in identifying and thinking through the correlation in the first place when the informational environment becomes complex enough. Only when the experimental instructions force subjects to focus on the correlation and the underlying independent signals, this

"nudge" strongly improves the rationality of belief formation. However, there may not always be an authority (like an experimenter) around the corner that helps overcome the biases due to correlation neglect. So, how can subjects' belief formation be debiased in other ways?

In this paper, we study whether decision making in teams is an effective way to avoid the biases resulting from correlation neglect, and in particular whether team composition matters. Economically relevant decisions are often made by teams, for which reason it seems fair to say that team decision is ubiquitous in both firms and governmental agencies, but also within households. In fact, there is by now a large body of evidence showing that team decisions can often be classified as more rational (in the sense of adhering more closely to the predictions of standard game theory under the assumption of selfish behavior) in comparison to individual decision making (for a review, see Kocher, Praxmarer, and Sutter, 2020, but also Charness and Sutter, 2012 and Kugler, Kausel, and Kocher, 2012). Yet, cognitive biases of teams, including correlation neglect, have not been studied in depth so far, with the exception of Fehrler, Hochleitner, and Janas (2026) which will be discussed below. Here we examine whether and to which extent teams can avoid correlation neglect by running a laboratory experiment with a total of 410 participants and how this might depend on the composition of teams.

We build our experimental design on Enke and Zimmermann (2019). As a benchmark against which to compare teams to, we also let individuals make choices and measure their degree of correlation neglect by comparing their estimates of the true state of the world in 10 estimation tasks in which they get either correlated or uncorrelated signals about the state of the world. In our team decision making treatment, we let two randomly paired subjects work on the same ten estimation tasks as in the individual treatment. Team composition is fixed for all tasks and team members can communicate via an electronic chat before making a joint decision. Importantly, in order to study team composition, we let all participants in our experiment do a single individual task *before* the ten estimation tasks. This allows us to classify each subject's level of naivete (in the sense of falling prey to correlation neglect) or rationality. From this measure, we can then classify whether a team consists of two naive members, a mix of one naive and one rational member, or two rational members. While team members are not informed about the other member's naivete or

rationality, we can identify whether the composition matters for the communication within teams – thus revealing the deliberation process – and for their decisions in the estimation tasks.

We establish four main findings. (1) Individuals show correlation neglect practically to the same degree as in Enke and Zimmermann (2019). Teams, on average, do not differ significantly from them, although the estimates are imprecise and do not rule out a sizable reduction. Suggestive evidence for the latter comes from the fact that teams earn significantly more money than individuals in the experimental task. (2) Team performance depends strongly on composition. The share of tasks in which a pair states the rational belief – rather than suffering from correlation neglect – rises from 9% with two naive members to 40% with one rational and one naive member and 67% with two rational members. (3) Teams with one rational member are, on average, about as naive as a rational individual alone, but they fall significantly short of a "truth wins" benchmark (Cooper and Kagel, 2005), under which a team states the rational belief whenever at least one member would state it alone (40% vs. 57% of tasks). (4) The chat data reveal why such teams falls short of the truth wins benchmark: they propose the rational solution about as often as a rational individual finds it alone, but adopt it in only about three quarters of these cases, because rational members cannot always convince naive members of the right approach to avoid correlation neglect.

Our paper brings together the literature on team decision-making (for a review, see Kocher, Praxmarer, and Sutter, 2020) and studies about errors in reasoning by individuals (e.g. Enke and Zimmermann, 2019; Enke, 2020; Charness, Oprea, and Yuksel, 2021; Moser and Wallmeier, 2021).[1] There is a rich literature on differences between decisions by teams and by individuals (Barber, Heath, and Odean, 2003; Cooper and Kagel, 2005; Kocher and Sutter, 2005; Kocher, Strauß, and Sutter, 2006; Maciejovsky and Budescu, 2007; Sutter, 2007; Luhan, Kocher, and Sutter, 2009; Adams and Ferreira, 2010; Keck, Diecidue, and Budescu, 2014; Buffat, Praxmarer,

1. Our paper also has a weak link to the literature on the accuracy of forecasting by experts and non-experts (e.g. DellaVigna and Pope, 2018; DellaVigna, Pope, and Vivalt, 2019). Empirically, DellaVigna and Pope (2018) find that individual experts are significantly less accurate than the average forecast, which speaks to an effect often called "wisdom of the crowds". In our paper, we do not compare individual choices to average choices of individuals, however, but examine whether teams are less prone to correlation neglect than individual decision makers, with a specific focus on the question whether team composition with respect to the naivete or rationality of its members make a difference.

and Sutter, 2023; Bauer et al., 2024). The main insight from these studies is that teams make more selfish decisions which moves their actions closer to standard game theoretic predictions under the assumption of selfishness. Much of this effect of team decision making has been ascribed to teams thinking through the decision problems more carefully by exchanging more information and with deeper levels of reasoning (see Charness and Sutter (2012) and Kugler, Kausel, and Kocher (2012) for the first reviews of this literature). The information sharing in teams helps avoid mistakes and make extreme decisions less likely (Adams and Ferreira, 2010). Perhaps surprisingly, there is not much work on the cognitive biases of team decision making. Houghton et al. (2000) note the persistence of biases (with regards to the law of small numbers and the illusion of control) in teams' risk perceptions compared to individuals. Mancuso et al. (2014) explore confirmation bias in team discussions. They find that when incorrect information is presented early, teams appear to focus more on that information in their subsequent discussions, and report incorrect answers more often. This finding is suggestive of confirmation bias in the team's deliberation process. Barahona et al. (2026) find that teams (of mutual fund managers) overreact less to recent investment performance than individuals, thus overshooting less. Most of this effect seems to be driven by the most biased team member influencing the team decision less than more unbiased members.

The paper most closely related to ours is by Fehrler, Hochleitner, and Janas (2026) and it was developed in parallel to ours (see page 3 of their forthcoming paper). They study correlation neglect of individuals and groups of three members each in a context where decision makers have to make binary choices, and receive signals – either correlated or uncorrelated ones – about the true state of the world. They find no evidence that groups perform better than individuals in a sense of being less prone to correlation neglect. As discussed by Fehrler, Hochleitner, and Janas (2026), theirs and our paper differ by the different correlation structure (which might affect results) and in the decision making process of teams. In their paper, decisions in groups are made by majority voting on proposals, while in our paper we request that teams of two members each reach a unanimous decision. So, in a sense our team decision making procedure is more likely to capture the spirit of team decisions in workplaces – where teams discuss pros and cons of different

options and often try to reach an agreement – whereas the group decision making procedure in Fehrler, Hochleitner, and Janas (2026) reflects more of a social choice setting in which voting determines a group outcome. We consider both approaches as complementary. Our paper is also different by focusing on the effects of team composition. Our design feature through which we can classify individuals as rather naive or rational allows us to examine how the different combinations of naive and rational members in teams affect decisions, and whether rational team members have positive externalities on naive members and how strongly the occurrence of correlation neglect depends on the number of rational team members. We supplement this by an LLM-based analysis of team chats and whether they depend on team composition as well.

The rest of the paper proceeds as follows: Section 2 describes the experimental design. Section 3 presents the results on correlation neglect. Section 4 examines the content of the team chats, and Section 5 concludes.

2 Experimental Design

We build our experimental design on Enke and Zimmermann (2019) and largely replicate their design, with extensions explained in detail below.

2.1 The Estimation Task - with correlated or uncorrelated signals

Participants are asked to estimate an *ex ante* unknown continuous state of the world (μ) and are compensated financially for accuracy (see Appendix D for the experimental instructions). The only information given to subjects are unbiased computer-generated signals (called "messages") about the true state. Participants are either shown (in a between-subjects design) correlated or uncorrelated signals about μ, as shown in Figure 1.

The fictitious computers A–D generate four unbiased, independent and identically distributed (i.i.d.) signals about μ, identical across participants and treatment groups (that are introduced

below). This is implemented by random draws from a truncated discretized normal distribution with mean μ and standard deviation $\sigma = \mu/2$. Truncation is at $\mu \pm 2\sigma = \mu \pm \mu$ in order to exclude negative signals. In the *Uncorrelated* condition (left panel, Figure 1), the intermediaries 1–3, which are also fictitious computers, observe the signals of computers B through D, respectively, and simply transmit these signals to the participant. This means that subjects receive information from both computer A and the intermediaries 1–3. For example, in one experimental task, the signals of computers A through D are given by 122, 90, 68 and 5, respectively.

In the *Correlated* condition (right panel, Figure 1), the three intermediaries observe the signal of computer A and computers B–D, respectively, and then report the average of these two signals. Participants are given information from computer A as well as from the intermediaries 1–3. In the example above, each of the three intermediaries takes the average of 122 (from computer A) and the corresponding signal of the other computer it communicates with. Therefore, computer A reports 122, intermediary 1 reports 106, intermediary 2 reports 95, and intermediary 3 reports 64. So, the information structure transmits messages to the participants that obscure the underlying signals. In our example, participants would see a majority of messages larger than 90, while the majority of signals was actually 90 or lower. Since the common signal from computer A is known, a rational participant would invert averages to back out the underlying independent signals from the messages of the intermediaries.

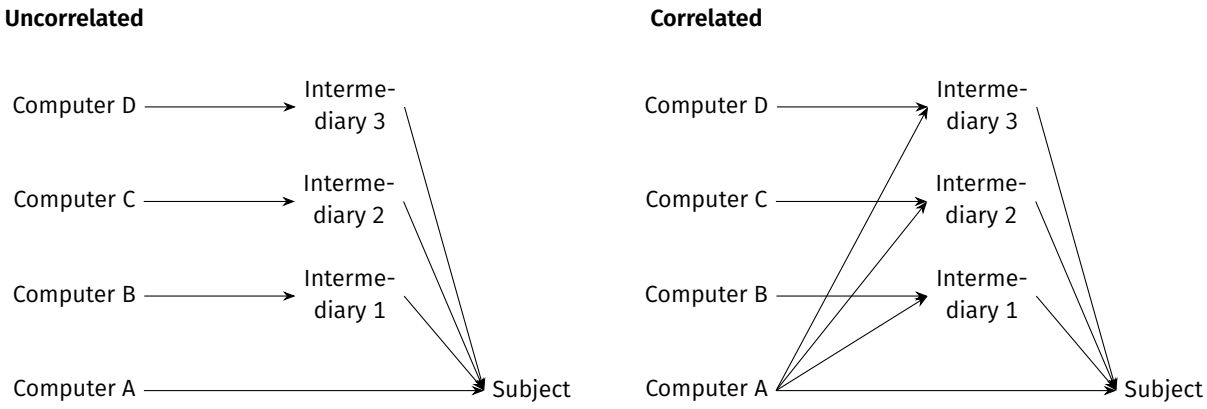


Figure 1. *Uncorrelated* (left panel) and *Correlated* (right panel) information structures

The identification strategy of Enke and Zimmermann (2019) depends entirely on the identical informational content of the two sets of signals. Differences in beliefs between the *Correlated* and *Uncorrelated* conditions can only be explained by variations in the information structure since all other factors are held constant. Thus, comparing beliefs across *Correlated* and *Uncorrelated* allows for the identification of correlation neglect. Participants are paid according to the accuracy of their stated belief in one randomly chosen task using a quadratic scoring rule (Selten, 1998), where variable earnings in euros are given by

$$\pi = \max\left[0,\ P - 160 \times \left(\frac{Belief}{True\ state} - 1\right)^2\right] \tag{1}$$

with P = 10.

2.2 Two parts of the experiment and payment rule

One important extension of our design in comparison to Enke and Zimmermann (2019) is to introduce a first part in which all participants have to individually give an estimate. This is called Part 1 in Table 1 and it is paid for sure according to a quadratic scoring rule, where earnings in euros are again given by Equation (1) with P = 2.

We will use the decisions in Part 1 in particular for classifying team composition with respect to the naivete or rationality of its members, as will become clear below. Part 2 of the experiment is a direct replication of the ten estimation tasks of Enke and Zimmermann (2019). Like them, we use three different randomized task orders (see Appendix C) and one randomly chosen task is paid out according to the formula given above. Note that participants do not get any feedback about potential payments of their estimates until all estimation tasks have been completed.

In Table 1 we show the correlated and uncorrelated signals for each estimation task in both parts of the experiment. It also shows the benchmarks of rational beliefs and "full correlation neglect" to indicate the direction and extent of a potential bias. Naive beliefs are defined as the average of the four messages that subjects receive in the *Correlated* condition, i.e., a posterior belief that fully falls prey to correlation neglect. In *Uncorrelated*, the task contains no correlation

to neglect: the rational belief is the simple average of the four messages. There, the classification of subjects into naive or rational captures accurate averaging rather than correlation sophistication.

Table 1. Overview of signal structures and benchmarks across rounds

	Intermediaries								
	Uncorrelated Signals			Correlated Signals			Benchmarks		
Computer A	**1**	**2**	**3**	**1**	**2**	**3**	**True**	**Rational**	**Naive**
Part 1									
390	460	680	810	425	535	600	500	585.00	487.50
Part 2									
12	9	10	0	10	11	6	10	7.75	9.75
122	90	68	5	106	95	64	88	71.25	96.75
179	295	288	277	237	234	228	250	259.75	219.50
565	847	650	1351	706	608	958	732	853.25	709.25
1110	1060	629	1100	1085	870	1105	1000	974.75	1042.50
1608	7240	4866	5526	4424	3237	3567	4698	4810.00	3209.00
9950	1203	11322	11943	5577	10636	10947	7338	8604.50	9277.50
2543	10780	6898	8708	6662	4721	5626	10000	7232.25	4888.00
15160	21806	20607	47751	18483	17884	31456	23112	26331.00	20745.75
12340	32168	49841	61293	22254	31091	36817	46422	38910.50	25625.50

Note: The reports of intermediaries 1–3 in the Uncorrelated condition directly reflect the draws of computers B–D. Signals from B–D are not directly observed but pass through intermediaries 1–3. There is no intermediary for computer A, which is directly observed. The rational belief benchmark is computed by taking the average of the signal of computer A and the reports of intermediaries 1–3 in the Uncorrelated condition. The naive belief benchmark is given by the average of the signal of computer A and the reports of intermediaries 1–3 in the Correlated condition. Participants faced the 10 tasks in one of three randomized orders. The same three orders were used in all treatments. Part 1 is our addition to the original design of Enke & Zimmermann (2019), shown in Part 2, and was completed individually by all participants. Part 2 was completed individually or as a team depending on the treatment assignment.

2.3 Treatments

We employ a 2 × 2 experimental design with four treatments, crossing the type of the decision maker (INDIVIDUAL vs. TEAM) with the information structure (*Correlated* vs. *Uncorrelated*). The INDIVIDUAL treatments are a replication of Enke and Zimmermann (2019), except that the time limit for submitting an estimate has been extended from five to seven minutes (in order to give teams enough time) and that Part 1 has been added.

In the TEAM treatments, participants perform in Part 2 in teams of two randomly matched subjects. This means that participants complete Part 1 as individual decision makers, but in Part 2 they need to make a joint decision about their team's estimation. Teams of two members remain fixed for the entire duration of Part 2. The signals that teams see are identical (also in the order in which they see them) to the INDIVIDUAL treatment. When they receive the signals, team members can make a proposal for an estimate via an electronic chat function, or they can start exchanging free-form, real-time messages before making a proposal. They have seven minutes to do so, and then can move on to the next estimation task if both team members endorse the same proposal. In case of a mismatch of proposals, they can continue chatting to find an agreement and enter new proposals within the time limit. In total, only in 2.7% of cases (36 out of 1,320) we observe that teams could not find an agreement.[2]

2.4 Hypotheses

INDIVIDUAL. In the information structure presented in Figure 1, the fictitious computers generate four i.i.d. signals $s_h \sim \mathcal{N}(\mu, (\mu/2)^2)$ (truncated at $(0,\ 2\mu)$) for $h \in \{1, \ldots, 4\}$. In the *Correlated* condition, participants are shown messages s_1 and $\widetilde{s}_h = (s_1 + s_h)/2$ for $h \in \{2,\ 3,\ 4\}$. When asked to estimate μ, a rational decision maker would back out the underlying independent signals from the messages $\widetilde{s}_h$ and calculate the mean rational belief as $b_B = \sum_{h=1}^{4} s_h/4$, which by design is also equal to the rational belief in the *Uncorrelated* condition.

Suppose the decision maker is naive and her stated posterior belief suffers from correlation neglect, i.e., she does not fully take into account the degree to which $\widetilde{s}_h$ reflects s_1, but instead treats $\widetilde{s}_h$ as independent. Following Enke and Zimmermann (2019), we label such a decision naive and let the degree of naivete be characterized by the parameter $\chi_{indiv} \in [0,\ 1]$ such that $\chi_{indiv} = 1$ implies full neglect, and $\chi_{indiv} = 0$ the fully rational belief. We classify a decision maker as rational if $|\chi_{indiv}| < 0.5$.

2. Those cases are excluded from all belief analyses below

A naive agent extracts s_h from $\widetilde{s}_h$ according to the following rule:

$$\widehat{s}_h \;=\; \chi_{indiv}\widetilde{s}_h + (1-\chi_{indiv})s_h \;=\; s_h + \tfrac{1}{2}\chi_{indiv}(s_1 - s_h) \tag{2}$$

where $\widehat{s}_h$ for $h \in \{2,3,4\}$ denotes the agent's (potentially biased) inference of s_h. She consequently forms mean beliefs according to

$$b_{CN,indiv} \;=\; \frac{s_1 + \sum_{h=2}^{4}\widehat{s}_h}{4} \;=\; \bar{s} + \tfrac{3}{8}\chi_{indiv}\left(s_1 - \bar{s}_{-1}\right) \tag{3}$$

where $\bar{s} = \left(\sum_{h=1}^{4} s_h\right)/4$ and $\bar{s}_{-1} = \left(\sum_{h=2}^{4} s_h\right)/3$. Therefore, a (potentially partially) naive belief is given by the rational belief $\bar{s}$ plus a belief bias component that depends on the naivete parameter χ_{indiv} and the magnitude of the common source signal (computer A) relative to the other three signals (computers B–D).

> ***Hypothesis 1.*** *Assuming that $\chi_{indiv} > 0$, beliefs in the Correlated condition exhibit an overshooting pattern. Given a high common source signal, i.e. $s_1 > \bar{s}_{-1}$, beliefs in the Correlated condition are biased upward compared to the Uncorrelated condition. Conversely, if $s_1 < \bar{s}_{-1}$, beliefs in the Correlated condition are biased downward.*

TEAM. Based on the predominant findings of the team decision making literature that teams are more rational agents (e.g., Charness and Sutter, 2012), we put forward the following hypothesis.

> ***Hypothesis 2.*** *Teams will form beliefs that are less biased toward the common source signal (of computer A) than those of individual decision makers. In other words, we expect a lower degree of naivete in team than in individual decisions.*

Teams form beliefs according to

$$b_{CN,team} \;=\; \frac{s_1 + \sum_{h=2}^{4}\widehat{s}_h}{4} \;=\; \bar{s} + \tfrac{3}{8}\chi_{team}\left(s_1 - \bar{s}_{-1}\right) \tag{4}$$

where $\widehat{s}_h$ is the team's inference of s_h, defined as in Equation (2) with χ_{team} in place of χ_{indiv}, and Hypothesis 2 predicts $\chi_{team} < \chi_{indiv}$ on average. It remains an open question whether this applies to all possible team compositions. From Part 1, we can classify participants as either naive or rational (with if $|\chi_{indiv}| = 0.5$ as the threshold). Due to randomly assigning teams of two subjects

each, teams can exist of (1) two rational members, (2) one rational and one naive members, or (3) two naive members. Partnering two rational subjects is unlikely to swing beliefs away from what one would expect from one rational individual, meaning that the distribution of estimates should not differ between rational subjects in the INDIVIDUAL treatment and teams with two rational members. A more interesting case is when teams consist of one naive and one rational member. The "truth wins norm" (Cooper and Kagel, 2005) would imply that rational agents can always convince the naive members to endorse a rational estimate. This would imply that mixed teams should perform as well as rational subjects in INDIVIDUAL and certainly better than naive subjects in INDIVIDUAL. Given that team members are *not* informed about the other member's type as being naive or rational, rational members may, however, not always be able to argue successfully for the rational solution. Finally, the type of teams with two naive members can be expected not to do better than naive subjects in INDIVIDUAL, and worse than rational subjects in INDIVIDUAL. This reasoning may only not apply in case the opportunity to discuss the task could enlighten naive members by finding the rational solution in the course of their discussion. Then teams of two naive members should outperform the naive subjects in INDIVIDUAL and may converge towards the rational subjects in INDIVIDUAL.

2.5 Procedural Details

The experiment was conducted at the Decision Lab of the Max Planck Institute for Behavioral Economics in Bonn, Germany. Subjects were mostly students from the University of Bonn and were recruited using the online recruitment system ORSEE (Greiner, 2015). No subject participated in more than one session. The experiment was run using the experimental software z-Tree v.4.2.6 (Fischbacher, 2007). Eight sessions were run in January–February 2020 and nine sessions in December 2022–January 2023. In total, 410 subjects participated.[3] For the INDIVIDUAL treatments, six sessions (19–29 subjects each) included 146 participants. For the TEAM treatments, 11

3. No subjects had to be excluded due to inadequate understanding of the task, which we checked with comprehension questions. Subjects could only proceed to the experimental task once all questions had been answered correctly.

sessions (14–30 participants each) yielded 132 teams (264 subjects). Each session was randomly designated as either INDIVIDUAL or TEAM, with participants then randomized to either the *Correlated* or *Uncorrelated* information structure. Sessions lasted approximately 1.5 hours. The average total payment was 16.87€ (SD 3.95, range 10–22€), consisting of a 5€ show-up fee, 5€ for the demographic questionnaire, and payments for Part 1 and Part 2 of the experiment. Table 2 shows that the characteristics of partipants were balanced across treatments (see the column with the p-values), except for earnings being significantly different between INDIVIDUAL and TEAM.

Table 2. Descriptive statistics by treatment

		INDIVIDUAL		TEAM		
Characteristic	**Overall**	**Correlated**	**Uncorrelated**	**Correlated**	**Uncorrelated**	**p-value**
Female (=1)	0.58 (0.49)	0.56 (0.50)	0.61 (0.49)	0.61 (0.49)	0.55 (0.50)	0.71
Age	23.15 (3.32)	23.18 (3.44)	22.69 (2.97)	23.12 (3.32)	23.40 (3.46)	0.74
Student (=1)	0.97 (0.16)	0.96 (0.20)	1.00 (0.00)	0.96 (0.20)	0.98 (0.15)	0.30
High-school grade	2.06 (0.64)	2.09 (0.67)	2.07 (0.61)	2.12 (0.62)	2.00 (0.67)	0.45
Avg. task earnings	5.71 (1.03)	5.09 (1.21)	5.73 (1.02)	5.59 (0.91)	6.12 (0.85)	<0.001
Number of subjects	**410**	**71**	**75**	**126**	**138**	

Note: Numbers show mean and standard deviation (in parentheses). Reported p-values represent results from Kruskal-Wallis rank-sum tests against the null hypothesis of equal distributions of the respective characteristic across the four treatment groups. Average task earnings is the profit derived from participants' beliefs in Part 2. Total payment also included a show-up fee, a questionnaire fee and the Part 1 payment, all paid on top of the amount shown.

3 Results on Correlation Neglect

3.1 Beliefs across *Correlated* and *Uncorrelated* Conditions

For INDIVIDUAL, Table 3 provides summary statistics across tasks and shows that in 7 out of 10 estimation tasks beliefs differ significantly at the 1% level between the *Uncorrelated* and the *Correlated* conditions, in the direction predicted by correlation neglect. This pattern confirms the seminal evidence by Enke and Zimmermann (2019). In all 10 tasks, the median belief in *Correlated*

is closer to the naive benchmark than this is the case in *Uncorrelated*. That holds true for small and large true states, and also for biases in either direction. This tendency of individuals to neglect the correlation is also reflected in average earnings for Part 2. As Table 2 shows, individuals in *Correlated* earn 5.09€ on average, which is 11.2% less than the 5.73€ in *Uncorrelated* ($p < 0.01$, Wilcoxon rank-sum test with one observation per decision maker).

For TEAM, Table 3 shows that beliefs differ significantly between *Uncorrelated* and *Correlated* in 3 out of 10 estimation tasks at the 1% level, and in 5 out of 10 at the 5% level. Teams show fewer significant task-level differences than individuals, but a count of significant tests does not show that teams are, in the aggregate, less prone to correlation neglect.[4]

However, teams earn significantly more money than individuals and thus outperform them clearly. Teams earn in *Correlated* 9.9% more than individuals in *Correlated* ($p < 0.05$). They also out-earn individuals in *Uncorrelated* (by 6.8%, $p < 0.01$). Thus, although teams are also prone to correlation neglect, they are on average closer to the rational estimates and earn significantly more money than individuals, a pattern that is in line with many other studies on team decision making (Kocher, Praxmarer, and Sutter, 2020). Yet, that teams suffer also from correlation neglect can be easily seen by the observation that they earn 8.6% less in *Correlated* than in *Uncorrelated* ($p < 0.01$). We can summarize these findings in our first result.

> ***Result 1:*** Both individuals and teams are prone to correlation neglect, even if this applies less to teams, on average. Regarding earnings from the task, teams outperform individuals significantly, however.

3.2 Heterogeneity and Team Composition

So far, we have found evidence of correlation neglect *on average*, for both individuals and teams. Averages may, however, mask considerable heterogeneity. For this reason, we proceed by estimating a naivete parameter χ for each individual and each team, where for the latter we consider a team's

4. As a descriptive complement, Tables A.1 and A.2 in the Appendix report task-level difference-in-differences analyses across conditions and within treatments using (normalized) absolute deviations from the rational benchmark.

Table 3. Correlation neglect by belief formation task and treatment

			Median Belief					
	Benchmarks		TEAM		INDIVIDUAL		p-value: Correlated vs. Uncorrelated	
True	Rational	Naive	Uncorrelated	Correlated	Uncorrelated	Correlated	TEAM	INDIVIDUAL
10	7.75	9.75	8.50	9.20	8.00	8.50	0.215	0.291
88	71.25	96.75	71.25	80.00	71.25	80.00	0.004	<0.001
250	259.75	219.50	260.00	260.00	260.00	250.00	0.177	0.001
732	853.25	709.25	845.00	800.00	850.00	780.00	0.273	0.007
1,000	974.75	1,042.50	1,000.00	1,010.00	994.25	1,007.30	0.815	0.103
4,698	4,810.00	3,209.00	4,810.00	4,810.00	4,810.00	4,500.00	0.034	<0.001
7,338	8,604.50	9,277.50	9,880.50	9,500.00	8,900.00	9,000.00	0.704	0.627
10,000	7,232.25	4,888.00	7,232.25	6,450.00	7,232.25	6,000.00	<0.001	<0.001
23,112	26,331.00	20,745.75	25,878.00	22,000.00	26,231.00	21,262.00	0.025	<0.001
46,422	38,910.50	25,625.50	38,910.50	38,000.00	38,915.00	33,350.00	0.004	<0.001

Note: This table reports median beliefs in the 10 estimation tasks by treatment and condition. The p-values refer to Wilcoxon rank-sum tests of equal belief distributions across the Correlated and Uncorrelated conditions within TEAM or INDIVIDUAL. For reference, we provide the true state and benchmarks for fully rational and fully naive beliefs. See Table 1 for details of the computation of the rational and naive benchmarks.

joint estimate. The model of belief formation in equations (2)–(4) links every stated belief to χ. For each belief b_i^j of decision maker i (an individual or a team) in task j, we solve equation (3) (individuals) or equation (4) (teams) for χ. This yields a task-level naivete value $\widetilde{b}_i^j$. Our estimate of a decision maker i's naivete is the median of these values across tasks:

$$\widehat{\chi}_i \equiv \text{med}\left(\widetilde{b}_i^j\right) = \text{med}\left(\frac{8\left(b_i^j - \bar{s}^j\right)}{3\left(s_1^j - \bar{s}_{-1}^j\right)}\right).$$

We exclude observations with $|\chi| > 3$, like Enke and Zimmermann (2019) did, who had about 4% such outliers. In our experiment, 3.4% of decision-maker × task observations (3.6% for individuals, 3.2% for teams) are classified as outliers by the same rule.

Table 4 presents the regression analysis examining the relationship between our treatments and correlation neglect, measured via naivete parameters. Complementary non-parametric tests for differences in distributions and ranks are provided in Table A.3 in the Appendix. We also use

χ^2-tests to compare the shares of subjects classified as rational versus naive based on their median naivete across tasks. We classify decision makers as fully rational/naive (both binary measures) if their median naivete (across tasks in Part 2) lies within ±0.2 of the benchmark of 0 (rational) or 1 (naive).[5]

We cluster standard errors by the decision maker. Column (I) shows a coefficient of 0.34 for *Correlated* in INDIVIDUAL, which is highly significant and very close to the estimates in Enke and Zimmermann (2019). Column (II) reveals that teams neglect correlation too, with a smaller coefficient though (0.20). Column (III) pools individuals and teams to test this difference. Under uncorrelated signals, teams do not differ significantly from individuals (TEAM: +0.03). The *Correlated* × TEAM interaction is −0.14 (SE 0.09, $p = 0.12$, 95% confidence interval from −0.32 to 0.04): teams shrink the *Correlated* gap by about 40%, but this difference is not statistically significant. Looking at the median naivete reveals a difference between individuals and teams, however. It is 0.30 for teams, but 0.55 for individuals in *Correlated*, a sizable difference that is weakly significant (Wilcoxon rank-sum test in Table A.3, $p = 0.06$). Teams as a whole thus appear less naive than individuals.

Figure 2 shows kernel density estimates and histograms of the distribution of naivete parameters for both the *Correlated* and the *Uncorrelated* conditions in the INDIVIDUAL and TEAM treatments. As expected, there is a density spike around zero for the *Uncorrelated* condition, indicating that the majority of subjects behave approximately rationally under the baseline scenario of no correlation in the information structure. This pattern holds for both the INDIVIDUAL and TEAM treatments. The *Correlated* condition, however, exhibits substantial heterogeneity. The distribution of naivete among individuals in *Correlated* has two main peaks, one around the rational benchmark ($\chi = 0$) and one around full correlation neglect (i.e. $\chi = 1$). Team beliefs in the *Correlated* condition exhibit a similar bimodal pattern to individuals, with a slight shift towards rationality.

5. Note that this classification for Part 2 is different from our approach for Part 1, where we defined rationality as implying $|\chi| < 0.5$, and naivete as the remaining cases. This was done in order to classify all subjects as naive or rational.

Table 4. OLS regressions: naivete and response time

	Naivete					
	Uncorrelated & Correlated				Correlated	
	IND	TEAM	All	IND	All	TEAM
	(I)	(II)	(III)	(IV)	(V)	(VI)
Correlated	0.344***	0.203***	0.342***	0.247***		
	(0.073)	(0.055)	(0.073)	(0.065)		
TEAM			0.026			
			(0.055)			
Correlated × TEAM			-0.142			
			(0.091)			
Pre-task naivete				0.248***	0.278***	
				(0.083)	(0.050)	
Response time (in min.)					-	
					0.063***	
					(0.020)	
Team Type 1						-
						0.382***
						(0.085)
Team Type 2						-
						0.609***
						(0.124)
Constant	0.171	0.588***	0.351***	0.140	0.618***	1.029***
	(0.166)	(0.109)	(0.079)	(0.138)	(0.194)	(0.147)
Num.Obs.	1408	1243	2651	1388	1266	578
R2 Adj.	0.074	0.163	0.104	0.139	0.139	0.237
Task FE	Yes	Yes	Yes	Yes	Yes	Yes
Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	No	Yes	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: OLS estimates with standard errors clustered at the decision-maker level (individual or team) in parentheses. One observation per decision maker and task: team members share the team belief, so each team enters once. The dependent variable in columns (I)–(VI) is the naivete implied by each of the 10 beliefs. The header rows give the sample of each column: conditions and treatments (IND = INDIVIDUAL, All = INDIVIDUAL and TEAM). Reference category for the team types: teams with two naive members (Team Type 0) in column (VI). Team Type = number of team members classified as rational in the pre-task (absolute pre-task naivete < 0.5). For teams, pre-task naivete is the mean of the two members' values. Session fixed effects are omitted in columns (III) because TEAM is assigned at the session level. All regressions exclude observations with absolute naivete > 3.

Result 2: Teams are slightly less naive than individual decision makers, but for both types of decision makers we observe a similar bimodal pattern, suggestive of rational and naive belief updating types.

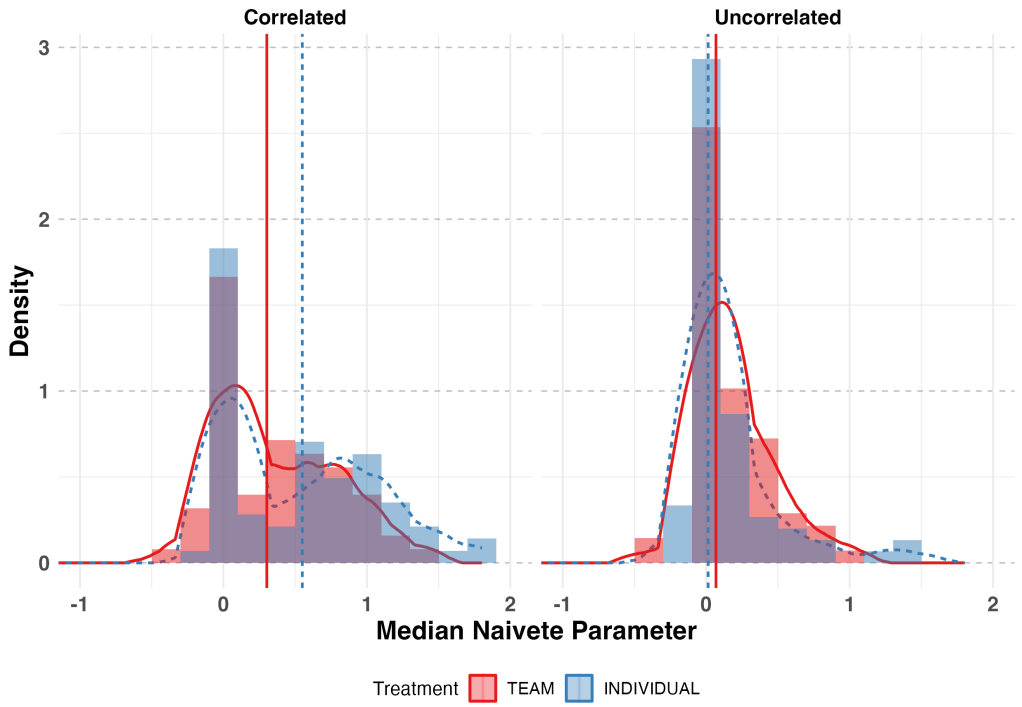


Figure 2. Heterogeneity in the median naivete parameter of beliefs across treatments

Notes: Kernel density estimates of each decision maker's median naivete, for individuals and teams, show the bimodal distribution of naivete in the *Correlated* condition and the unimodal distribution in the *Uncorrelated* condition. Colored vertical lines indicate medians within the respective distribution. Benchmarks for perfectly rational (naive) beliefs are at 0 (1).

Next, we examine team composition according to our classification specified above in Section 2.4. Using the naivete measure from Part 1, subjects are classified as rational if the absolute value of naivete is below 0.5 and naive otherwise. Importantly, the distribution of naivete in Part 1 does not differ between subjects in INDIVIDUAL and TEAM within either condition (Wilcoxon rank-sum tests, $p = 0.54$ and $p = 0.38$). Table A.4 in the Appendix shows the percentage of individuals classified as rational or naive according to this Part 1 assessment. Given that teams have been paired randomly, they can be of three different types: Type 0 teams have zero rational members, but two naive ones, Type 1 teams are composed of one naive and one rational member (mixed teams), and Type 2 teams have two rational members. In *Correlated*, we have 20 Type 0-teams, 35 Type 1-teams, and 7 Type 2-teams. The numbers for *Uncorrelated* are 3, 27, and 37.

Before diving deeper into the effects of team composition, we need to establish first that there is a relationship between behavior in Part 1 and the estimates in Part 2 of our experiment. In fact, the classification as naive or rational separates types sharply: Looking only at individual Part 1-types in INDIVIDUAL and *Correlated*, we see that the median naivete in Part 2 is 0.82 for those classified as naive in Part 1 and 0.01 for those classified as rational. Hence, the classification in Part 1 is a useful proxy for later behavior. Column (IV) in Table 4 confirms this on an even broader basis by considering both *Correlated* and *Uncorrelated* and showing that, in INDIVIDUAL, a subject who is fully naive in Part 1 has a naivete parameter in Part 2 that is 0.25 higher than for a fully rational subject ($p < 0.01$). Column (V) reports a similar coefficient of 0.28 ($p < 0.01$), while also controlling for response times and using only data from *Correlated*.

Column (VI) of Table 4 restricts the sample to teams in *Correlated*. Compared with Type 0-teams, naivete is 0.38 lower in Type 1-teams and 0.61 lower in Type 2-teams, both $p < 0.01$. The further step from one to two rational members (−0.23) is marginally significant ($p = 0.07$). Because partners are matched randomly after Part 1, a member's partner type is unrelated to the member's own type by design. The contrast between Type 1- and Type 0-teams therefore measures for a naive member the effect of having a rational rather than a naive partner, and the contrast between Type 2- and Type 1-teams the same effect for a rational member.[6]

This heterogeneity due to team composition is apparent in the density plots of Figure 3, which show very clearly that teams with two naive members are more likely to suffer from correlation neglect than teams with at least one rational member (Wilcoxon rank-sum test, $p < 0.01$). For a naive member, a rational partner thus lowers the team's naivete, on average, by about half (0.38 against a mean of 0.72 in Type 0-teams). In medians, the reduction is much larger, from 0.72 to 0.05. Both summaries are sensitive to the bimodal distribution of naivete (Figure 3): most beliefs

6. Subjects classified as rational in Part 1 have better high-school and math grades than naive subjects (1.94 vs. 2.24 and 2.14 vs. 2.58 on the German scale with 1 as the best grade, both $p < 0.001$ in Wilcoxon rank-sum tests) and are less often female (52% vs. 66%, $p = 0.004$). They do not differ significantly in age, student status, income or laboratory experience. Among teams under correlated signals, these differences largely vanish. Teams with one and with no rational member do not differ significantly in any of these characteristics (tests on team averages, all $p \geq 0.22$). Teams with two rational members have somewhat better high-school grades than teams with one rational member (1.76 vs. 2.15, $p = 0.06$) but not better math grades (2.21 vs. 2.34, $p = 0.81$). This may reflect differences in general effort or motivation rather than in numerical skills.

lie near either the rational or the naive benchmark, so the median sits at whichever mode holds more of the beliefs, while the mean lies between the modes and moves with the share of beliefs at each.

But is a mixed team as good as its rational member would be alone? On naivete, the answer is roughly yes: in *Correlated*, Type 1-teams are not significantly more naive than rational individuals (median naivete 0.05 vs. 0.01). A stricter criterion is whether a team states the rational belief, which we define as a belief whose implied naivete lies within ±0.2 of the rational benchmark ($|\chi| \leq 0.2$). The share of tasks in which a team states the rational belief rises from 9.1% in Type 0- to 39.8% in Type 1- and 67.2% in Type 2-teams (+30.6pp, $p < 0.01$, and +27.4pp, $p < 0.05$). Table 5 compares these shares with benchmarks built from individuals in *Correlated*, grouped by their own Part 1 type. The simplest benchmark is the team's better member alone (column *Better member*): rational individuals state the rational belief in 51.0% of tasks. The difference from Type 1 teams of −11.2 percentage points (pp) is sizable but imprecisely estimated (95% confidence interval from −26.2 to 3.8pp, $p = 0.14$). This comparison needs no assumption about how members interact. With a band of ±5% around the rational benchmark instead, the difference is larger and significant (Appendix A.2).

Comparing a mixed team with a rational individual asks whether the team loses anything by having one naive member, and the answer is mainly no. A more demanding question is whether the team gains from having two members, even if one may be naive. The "truth wins" model of group problem solving (Lorge and Solomon, 1955; Cooper and Kagel, 2005) provides a benchmark for this: it assumes that a team states the correct answer whenever at least one member would find it alone, because the member who has the answer convinces the other. Under this rule, a pair does at least as well as its stronger member, and better whenever the weaker member gets a task right that the stronger one gets wrong. We do not observe what team members would have answered alone, so we build the benchmark from individuals of the same Part 1 type in INDIVIDUAL ("simulated teams" in Table 5). Within each task, we pair every naive with every rational individual in *Correlated* and count a pair as successful if at least one of its members states

the rational belief. Such virtual pairs succeed in 57.2% of tasks.[7] The benchmark is modest because rational individuals themselves state the rational belief in only about half of the tasks (51.0%), and a naive partner, who does so in 13.5% of tasks, adds about six percentage points.

Type 1-teams state the rational belief in 39.8% of tasks. Table 5 places this figure among three benchmarks for a mixed pair. The least demanding is to compare Type 1-teams to a simulated pair that splits the difference and states the mean of its two beliefs (*Averaging*). Such pairs would state the rational belief in only 18.2% of tasks. Type 1-teams do significantly better than averaging (+21.6pp, $p < 0.01$). A medium benchmark is to compare Type 1-teams to the frequency with which rational individuals in Part 1 would choose the rational estimate in Part 2. That would be in 51.0% of cases (see column *Better member*). Type 1-teams are not significantly different from this benchmark. Yet, they fail the most demanding benchmark, which builds on the truth-wins approach, according to which simulated pairs would state the rational estimate in 57.2% of tasks. Type 1-teams do 17.4pp worse ($p < 0.05$, see column *Trust wins*). Most of this shortfall is not a failure to use the naive member's occasional correct answers, which are worth only 6.2 percentage points (57.2% vs. 51.0%). It arises because Type 1 teams state the rational belief less often than their rational member would alone (−11.2pp), although this difference is not significant on its own.

The other team types fit into this pattern. Teams with two naive members (Type 0 teams) do no better than one naive individual alone (9.1% vs. 13.5%, −4.4pp, n.s.), so a second naive member does not improve rationality. They fall significantly short of the truth-wins benchmark (−16.2pp, $p < 0.05$), and in naivete, too, they resemble naive individuals (median 0.72 vs. 0.82; n.s.).

Teams with two rational members (Type 2 teams) are the only type that does better than one member of its type alone (67.2% vs. 51.0% for a rational individual, +16.2pp), and they come

7. A truth-wins formula gives almost the same value. A pair fails only if both members fail. If member i fails with probability $1-p_i$, member j with probability $1-p_j$, and the two fail independently of each other, both fail with probability $(1-p_i)(1-p_j)$, so the pair succeeds with probability $1-(1-p_i)(1-p_j)$. With the rates of naive ($p_i = 13.5\%$) and rational individuals ($p_j = 51.0\%$), this is $1-0.865 \times 0.490 = 57.6\%$. Independence fails if some tasks are harder than others for everyone: both members then tend to fail on the same tasks, and the formula overstates the benchmark. Pairing individuals within each task accounts for this. Since the two values nearly coincide (57.2% vs. 57.6%), differences in difficulty across tasks matter little.

closest to their truth-wins benchmark (−7.9pp). Neither difference is significant, though, and with seven Type 2 teams these estimates are imprecise.

Table 5. Rational beliefs of teams and simulated teams in the Correlated condition

		Simulated team (benchmark)		
	Team	**Averaging**	**Better member**	**Truth wins**
Type 0 (two naive)	9.1	8.5	13.5	25.3
Team minus benchmark (pp)		+0.6	-4.4	-16.2**
		(3.6)	(5.0)	(7.4)
Type 1 (one naive, one rational)	39.8	18.2	51.0	57.2
Team minus benchmark (pp)		+21.6***	-11.2	-17.4**
		(5.8)	(7.6)	(7.3)
Type 2 (two rational)	67.2	43.7	51.0	75.0
Team minus benchmark (pp)		+23.5*	+16.2	-7.9
		(12.5)	(12.5)	(12.4)

Note: Correlated condition, one observation per decision maker (individual or team) and task. Share of tasks in which the decision maker states the rational belief (final belief with absolute naivete of at most 0.2), in percent. Rounds without agreement are excluded. Individuals are classified by their Part 1 naivete (rational if absolute pre-task naivete < 0.5). Team type indicates the number of rational members. Simulated teams: within each task, all virtual pairs of individuals from the INDIVIDUAL treatment with the composition of the team type. The pair states the mean of the two beliefs (Averaging), the belief of the member with the better Part 1 type (Better member, the rate of individuals of that type), or the rational belief if either member does (Truth wins). Rates average over pairs within a task, then over tasks. Cluster-bootstrap standard errors in parentheses (2000 replicates, with individuals resampled within type and teams within team type). * p < 0.1, ** p < 0.05, *** p < 0.01.

Result 3: Under correlated signals, teams with one or two rational members are significantly less naive than teams with two naive members, and each rational member significantly raises the share of tasks in which the team states the rational estimate (9%, 40% and 67%). With one rational member, teams are on average about as naive as a rational individual would be alone but state the rational belief significantly less often than the truth-wins benchmark would predict (40% vs. 57%).

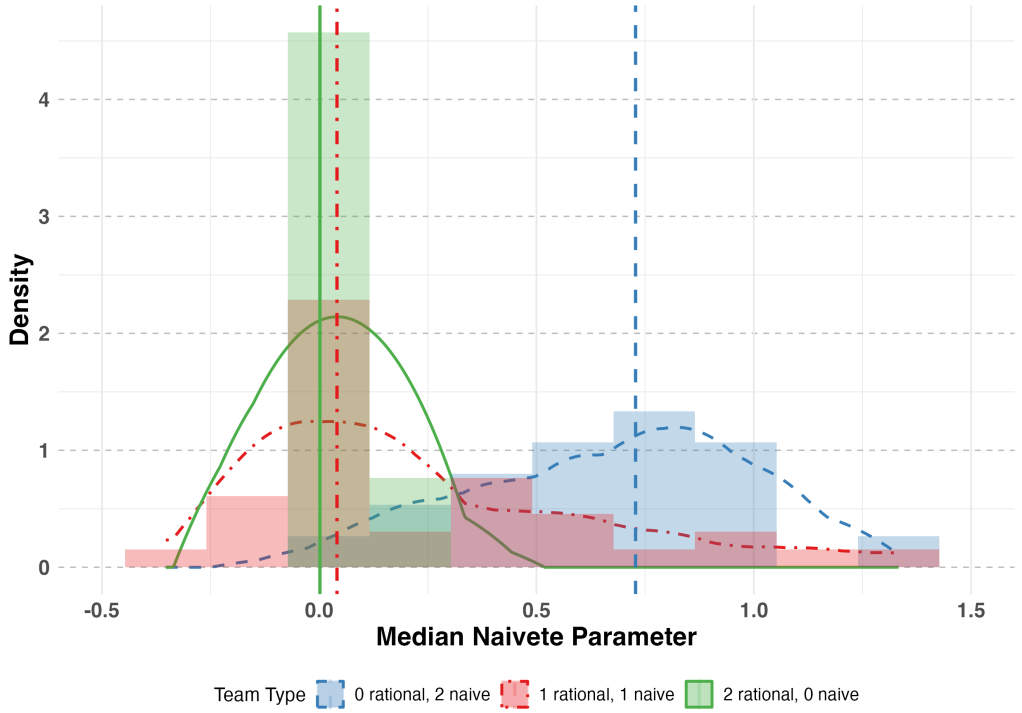


Figure 3. Heterogeneity in median naivete parameters of teams in the *Correlated* condition

Notes: The three kernel density plots show the distributions of naivete in teams of varying composition of rational and naive members. Before the 10 joint tasks, each subject completed an individual belief formation task, and the implied naivete of the belief was computed. Team members are classified as rational if the absolute value of their pre-task naivete is below 0.5 and as naive otherwise. The team type is the number of rational members (0, 1 or 2).

4 Examining the team chat content

In this section, we analyze the content of the communication within teams in order to learn about the decision making process and, in particular, how different team composition affects the results. We begin by explaining the methodology applied to code the chat content, which is based on using large language models (LLMs) and their validation through human coders. After that, we present the results of the chat analysis.

4.1 Coding the Team Chat Content

The chat is the complete record of how a team reached its joint estimation. We code each of the 1,320 team-round transcripts from Part 2 (132 teams × 10 tasks, each typically 3–15 messages in

informal German) in two ways: a reasoning-strategy category assigned by LLMs, and the rationality of the proposed solution, extracted from the numbers in the transcript. Appendix B documents every step. In the following, we summarize the main steps in this chat classification.

Coding scheme. We use the corpus of a team chat as the source for classifying categories of reasoning with which team members try to explain or justify their proposed estimates. For this purpose, we let four LLMs from different model families propose mutually exclusive reasoning categories. This was done by having them analyze 50 randomly selected transcripts, and each LLM did this ten times (with 50 newly randomly sampled transcripts) (see the parameters in Table B.1). Categories that did not recur across these ten subsamples were dropped. Then, codebooks from the four LLMs were merged into 15 categories. Nine of them were proposed by at least two model families. Table B.2 in the Appendix shows the codebook and each category, including a label, a definition and an example.

Human reference coding. Two human coders, blind to the hypotheses and trained separately, independently coded a subsample of 10% of the total chat transcripts. Coders were asked to assign each message to a primary and secondary category, and also to indicate how confident they were in their assignment. Given that the codebook had 15 categories, it is unavoidable to also have diverging assignment across both coders. Against this background, the 39.4% of cases in which both coders agreed on the same primary category is encouraging. The other messages were assigned according to the following rule: if both coders listed a particular category as primary or secondary, a message was assigned to that category (18.2% of cases), otherwise the primary category of the relatively more confident coder was used (42.4% of cases). Disagreements were not discussed. This approach is in line with Haaland et al. (2025) who note that where objectivity is the goal, fewer interactions between coders are preferable, because discussion can produce interpretative convergence.

Classification by LLMs. Four open-weight models (Mistral Medium 3.5, Gemma 4, Llama 3.3 and Qwen 3.6) then classified all transcripts. They ran locally, so no chat data left the institutional

cluster. Each model received the verbatim codebook and the same instruction as the human coders. Decoding was deterministic and the output structured.

Validation. As already indicated, the two human coders agreed on the primary category in 39.4% of transcripts (Cohen's $\kappa = 0.32$). By conventional thresholds this may look low, but it is typical for a 15-way judgment on open-ended reasoning: reliability falls with the number of codes and with conceptual difficulty (Haaland et al., 2025). We therefore benchmark the models against this human ceiling, scoring each model against each coder separately, as shown in Table 6. The better models agree with the coders about as well as the coders agree with each other: Qwen 3.6 reaches $\kappa = 0.35$ against coder A and 0.29 against coder B, Gemma 0.38 and 0.27, Mistral 0.37 and 0.23, and Llama 0.29 and 0.23. Against the human gold standard, which combines both coders' labels, the models reach κ between 0.29 and 0.35. The claim is therefore not that the machine coding is accurate, but that the better models agree with each human coder about as well as the two human coders agreed with each other.

Table 6. Agreement of the machine coding with each human coder and with the human gold standard

	Cohen's κ against			Against the human gold standard			
Coder	Coder A	Coder B	Human gold	Strict acc.	Partial acc.	Krippendorff's α	Gwet's AC_1
Human coders							
Human coder A	—	0.323	0.625	0.682	—	0.625	0.660
Human coder B	0.323	—	0.694	0.705	—	0.694	0.713
Models							
Gemma 4 31B	0.379	0.274	0.346	0.432	0.674	0.342	0.392
Qwen 3.6 27B	0.351	0.293	0.344	0.432	0.667	0.342	0.392
Mistral Medium 3.5	0.374	0.230	0.313	0.409	0.659	0.312	0.367
Llama 3.3 70B	0.293	0.228	0.288	0.386	0.629	0.285	0.343

Agreement of each coder's primary category on the same 132 double-coded transcripts. Cohen's kappa against coder A and coder B: agreement with each human coder separately. In the human rows, this is the agreement of the two coders with each other. Human gold: the reference standard built from both coders' labels. Because it is built from their own labels, the human coders' agreement with it is high by construction and cannot be compared with the models'. Strict accuracy: share of transcripts on which the primary category equals the gold label. Partial accuracy: share on which the model's primary category matches any label that either human coder flagged as primary or secondary (undefined for the human rows). Krippendorff's alpha and Gwet's AC1: chance-corrected agreement with the gold label over the 15-category nominal scale.

Choice of coder LLM. The four models cannot be ranked: every pairwise McNemar test on the validation sample gives $p \geq 0.38$. We use Qwen 3.6, because it has the highest accuracy on the transcripts on which both human coders chose the same category: 0.615, against 0.577 for Mistral, 0.558 for Gemma and 0.500 for Llama. On the transcripts on which the human coders disagreed and the reference label rests on the confidence tiebreak, it does no better than the other models (0.268, against 0.232 to 0.304). Since no difference in accuracy is significant, the choice is a convention rather than a ranking, and the appendix repeats every category result under all four coders. The four models never disagree on the direction of a difference of at least one percentage point between the *Correlated* and *Uncorrelated* conditions (Table B.3). Table B.4 reports every coder, conditional by team type (where type x means to have x rational members). Individual category gaps can differ in size across models, so we interpret only differences on which the coders agree.

Solution indicator regarding the estimate. Whether a rational estimate was proposed is a factual question, which we answer without an LLM. Rather, we extract every number from the transcripts, discard numbers that equal one of the four computer signals or an intermediary's message under either information structure, or a sum of them, and flag the round if any remaining number is the rational estimate: the latter is defined as a number whose implied naivete lies within ± 0.2 of the rational benchmark ($|\chi| \leq 0.2$, as defined in Section 3.2).[8]

4.2 Analysis of the Team Chat Content

We start the analysis by, first, showing how team members communicate with each other. More specifically, we present in Table 7 the relative frequency of reasoning categories our four LLMs have identified in the team chats. These categories refer to how teams discuss their different proposals and how they try to find an agreement on a specific proposal. In the table, we show only data for *Correlated*, as this is the most interesting treatment to study correlation neglect. However, in

8. As a robustness check, we report in Appendix A.2 the results if we define a rational estimate to lie within a $\pm 5\%$ band of the rational solution rather than via $|\chi| \leq 0.2$. The results remain qualitatively the same.

Table B.3 in the Appendix, we show that the category distributions differ only modestly between *Correlated* and *Uncorrelated*.[9]

For Type 0- and Type 1-teams, we see in Table 7 that direct agreement – after one member has made a proposal –, intuitive guesses and compromises on the midpoint of both members' proposals are the most frequent ways of how such teams reach an agreement, covering more than 50% of the reasoning categories that our LLMs classify. Comparing the relative frequencies of the 15 categories across Type 0- and Type 1-teams reveals no significant differences. So, the way how these teams reason about the best estimate is not different, but we know from the previous section that they make significantly different estimates. This implies that the generation of (rational) proposals must differ, and we will return to this issue below.

For Type 2-teams, the reasoning pattern in the conversations looks different from the other team types. The prevalence of the share of intuitive guesses falls from 20.3% to 5.7% and that of a midpoint compromise from 14.9% to 2.9%, while the iterative refinement of proposals rises from 4.3% to 25.7%. All three shifts are significant at $p < 0.05$ after Holm correction. Their direction is the same under all four LLMs, but their size varies, and iterative refinement is significant only for the QWEN and for all LLMs' average (see Table B.4).

> ***Result 4a:*** Having one rational member in a team does not change the way team members communicate with each other compared to teams with two naive members. The mix of reasoning categories does not change significantly. If teams are composed of two rational members, however, chats are coded less often as guessing or midpoint compromises and more often as iterative refinement of proposals.

So, what happens in Type 1-teams compared to Type 0-teams, regarding their estimates? Section 3.2 had already shown that one rational team member lowers a naive member's naivete by about one half, which speaks in favor of less correlation neglect in teams with one rational member. However, in that section we had also seen that Type 1-teams do worse than an ambitious truth-wins benchmark predicts (39.8% vs. 57.2% rational estimates). Our chat data can help us

9. Under correlated signals teams guess more (Intuitive Guess +6.4pp) and adjust for outliers less often (−6.0pp), and all four LLMs agree on both directions. The difference in adjusting for outliers is partly mechanical: in *Uncorrelated* the intermediaries pass on raw signals, so extreme draws are visible, whereas in *Correlated* each report averages two signals. No category gap survives Holm correction for the reported coder, so these differences are descriptive.

Table 7. Coded communication by team type, Correlated condition

	Share by team type (%)			Difference (pp)		
Category	**Type 0**	**Type 1**	**Type 2**	**1 vs 0**	**2 vs 0**	**2 vs 1**
Direct Agreement	24.0	19.1	21.4	-4.9	-2.6	+2.3
Intuitive Guess	26.5	20.3	5.7	-6.2	-20.8**	-14.6**
Midpoint Compromise	17.5	14.9	2.9	-2.6	-14.6***	-12.0**
Direct Calculation	9.5	12.3	20.0	+2.8	+10.5	+7.7
Outlier Adjustment	7.0	13.4	8.6	+6.4	+1.6	-4.9
Rounding & Simplification	6.5	7.7	4.3	+1.2	-2.2	-3.4
Iterative Refinement	4.5	4.3	25.7	-0.2	+21.2***	+21.4***
Humor & Banter	1.5	2.9	2.9	+1.4	+1.4	+0.0
Algorithmic Correction	0.0	2.6	2.9	+2.6	+2.9	+0.3
Procedural Coordination	0.5	1.4	1.4	+0.9	+0.9	+0.0
Sum Verification	0.0	0.6	2.9	+0.6	+2.9	+2.3
Benchmark Anchoring	0.0	0.6	0.0	+0.6	+0.0	-0.6
Expressed Uncertainty	0.5	0.0	1.4	-0.5	+0.9	+1.4
Social Small Talk	1.0	0.0	0.0	-1.0	-1.0	+0.0
Weighted Calculation	1.0	0.0	0.0	-1.0	-1.0	+0.0

Note: Columns 2–4 give the category share for each team type in percent. Columns 5–7 give the pairwise differences in percentage points. Teams per type: Type 0 = 20, Type 1 = 35, Type 2 = 7. Cluster bootstrap over teams, 2000 replicates. *, ** and *** mark p < 0.1, 0.05 and 0.01 after Holm correction across the 15 categories within a contrast. Differences where one of the two team types never shows the category are not tested, because the bootstrap distribution is then degenerate.

better understand what's going on. More precisely, we can examine two steps that are necessary for a team to submit a rational estimate (rather than a naive one that suffers from correlation neglect): (1) *generation*, i.e., whether the rational answer (a number with $|\chi| \leq 0.2$) is proposed in the chat in the first place, and (2) *adoption*, i.e., whether a team that proposed a rational solution also submits it as its joint estimate. We can study the generation of rational proposals by accessing the solution indicators described in Section 4.1 and by retrieving the joint estimate from the experimental software. So, here we don't rely on the LLM coding. However, we use the LLM categories to understand whether and how teams convert proposals into joint estimates. For that purpose, we can use the 129 classified teams (62 in *Correlated*). The *p*-values come from a cluster bootstrap over teams (2,000 replicates).

Generation. Table 8 reports both steps by team type under correlated (Panel A) and uncorrelated signals (Panel B), together with the differences between team types. Under correlated signals, the rational answer is proposed in 18.0% of rounds in teams with no rational member (Type 0), in 53.4% with one (Type 1) and 82.9% with two (Type 2). Both increases are significant (+35.4 and +29.4pp, both $p < 0.01$). Under uncorrelated signals the gradient is much flatter (56.7%, 69.3% and 72.2%, both contrasts are not significant). The Part 1-classification means something different in *Uncorrelated*, however: it captures accurate averaging rather than correlation sophistication. A significant gradient that appears only under correlated signals is therefore hard to reconcile with general ability alone.

Table 8. Generation and adoption of the rational answer in the team chat

	By team type			Difference between types	
	Type 0	**Type 1**	**Type 2**	**Type 1 − 0**	**Type 2 − 1**
Panel A: Correlated					
Generation	18.0	53.4	82.9	+35.4*** (6.5)	+29.4*** (8.1)
Adoption	50.0	75.0	81.8	+25.0** (11.8)	+6.8 (12.1)
Final belief rational	9.1	39.8	67.2	+30.6*** (5.9)	+27.4** (12.1)
Panel B: Uncorrelated					
Generation	56.7	69.3	72.2	+12.6 (20.1)	+2.9 (6.9)
Adoption	70.6	75.0	70.5	+4.4 (15.1)	-4.5 (7.7)
Final belief rational	43.3	51.9	50.7	+8.5 (14.0)	-1.2 (8.4)

Note: One observation per team and round. Teams are classified by the number of members who were rational in the Part 1 task (team type). The rational answer is a number whose implied naivete lies within +/-0.2 of the rational benchmark. Generation: share of team-rounds in which the rational answer is mentioned in the chat (all numbers are extracted from the transcript, excluding numbers equal to a computer signal, an intermediary message under either information structure, or a sum of these). Adoption: share of these rounds in which the team's final belief is the rational answer. Final belief rational: share of all rounds in which the final belief is the rational answer. Adoption and final belief exclude rounds without agreement. Shares in percent by team type. Differences between adjacent team types in percentage points. Cluster-bootstrap standard errors in parentheses (teams resampled with replacement within condition and team type, 2000 replicates). *, ** and *** mark p < 0.1, 0.05 and 0.01 (normal approximation).

Result 4b: Team composition matters a lot for whether rational proposals are made or not. With correlated signals, a rational estimate is proposed in 18%, 53% and 83% of cases by teams with zero, one and two rational members. There is no such significant gradient with uncorrelated signals.

Adoption. Once a rational estimate is voiced in a team and particular task, teams with more rational members are also more likely to implement it as their team decision: 50.0%, 75.0% and 81.8% of such tasks are concluded with a rational joint estimate (Type 1 versus Type 0 +25.0pp, $p < 0.05$, and Type 2 versus Type 1 +6.8pp, n.s.).

Taken together, these two steps decompose the shortfall relative to the truth-wins benchmark found in Section 3.2. Mixed teams propose the rational answer about as often as a rational individual reaches it alone (53.4% vs. 51.0%). Yet, they adopt it in only about three quarters of these cases, so that they finally submit a rational belief in 39.8% of cases. The truth is voiced, but it does not always get implemented. Mixed teams under uncorrelated signals, by the way, implement a proposed rational estimate about equally often (in 75.0% cases), so the ”loss" of about one quarter of rational estimates is not specific to correlated information.

> ***Result 4c:*** Once proposed, the rational answer becomes the team's joint estimate in 50%, 75% and 82% of cases with zero, one and two rational members. Mixed teams propose it about as often as rational individuals find it alone, but adopt it only about three quarters of the time, which explains why mixed teams fall short of a truth wins benchmark.

5 Conclusion

Previous research has shown that many individuals neglect correlation between information sources, thus creating systematic biases in decision making (Enke and Zimmermann, 2019). We can replicate this pattern of individual decision making in our experiment, but ask the main question whether team decision-making reduces the bias from correlation neglect. In particular, we are interested in whether team composition - with respect to the number of naive or rational members - matters for team decisions and whether we can use the communication in teams to shed light on what's going on in teams.

At first sight, teams are, on average, no better than individuals with respect to correlation neglect. However, this first impression becomes more nuanced when observing that median choices

by groups are weakly significantly less affected by correlation neglect and that teams earn significantly higher amounts of money in our experimental task (in which earnings depend on the degree of correlation neglect). So, by and large, our findings support the main takeaway from the economics literature on team decision making that teams often make more rational choices (Charness and Sutter, 2012; Kugler, Kausel, and Kocher, 2012; Kocher, Praxmarer, and Sutter, 2020).

Hidden beneath these aggregate results, we discover large heterogeneity with respect to the composition of teams. Based on an individual decision making task before the main experiment, we have been able to classify individuals as naive (with respect to correlation neglect) or rational. Given that teams of two members each were randomly assigned, we can classify different types of teams, depending upon whether they have zero rational member (Type 0), one rational member (Type 1), or two rational members (Type 2). The share of tasks in which a team states the rational estimate and thus avoids correlation neglect rises from 9% with two naive members (Type 0-teams) to 40% with one rational member (Type 1) and 67% with two (Type 2). Pairing one naive member with another naive member has no benefits, meaning that two heads are not better than one in such a team composition. Yet, pairing a naive member with a rational one reduces the naive member's naivete by about one half. Moreover, mixed teams of one rational and one naive member make a rational estimate about as often as rational individuals alone. Yet, they fall short of a demanding "truth wins" benchmark according to which a team would be expected to choose the rational estimate whenever at least one member would choose it alone. There are only a few studies where teams beat such a benchmark (see Cooper and Kagel (2005) as one example), so our results are in line with the large majority of studies as mixed teams choose the rational estimate in 40% of cases, when the "truth wins" benchmark would be 57%. Our analysis of the team chats shows where the shortfall arises. Mixed teams propose the rational estimate about as often as a rational individual finds it alone, but adopt it in only about three quarters of these cases. The rest of the time they settle on another number, often compromising on an average with the naive member's proposal. The mix of coded reasoning categories does not differ significantly between teams with zero and one rational member, but it is the frequency of making rational proposals that makes the difference. Having a second rational member in the team improves the likelihood

of rational estimates further, and it also changes the most frequent categories of discussion. With two rational members, they make less intuitive guesses and compromise less often on midpoints between both members' proposals. Instead, they more often work on iterative refinements of their choices.

For practice, the results suggest that the benefit of deciding in teams depends on who is in the team. Under correlated signals, teams tend to be less naive than individuals, though not significantly so at the 5% level (median naivete 0.30 vs. 0.55, $p = 0.06$). In particular, the gain is concentrated among naive members who have a rational partner, as this reduces naivete by about one half, implying that correlation neglect is less of a problem for teams than for individuals. This speaks in favor of team decision making, which is, in fact, a ubiquitous practice in most organizations.

References

Adams, Renée B., and Daniel Ferreira. 2010. "Moderation in Groups: Evidence from Betting on Ice Break-ups in Alaska." *The Review of Economic Studies* 77 (3): 882–913. [5, 6]

Akerlof, George A., and Robert J. Shiller. 2009. *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism.* Princeton, NJ: Princeton University Press. [3]

Barahona, Ricardo, Stefano Cassella, Kristy A. E. Jansen, and Vincenzo Pezone. 2026. "Do Teams Alleviate or Exacerbate Overreaction in Beliefs?" *Journal of Financial Economics* 176: 104219. [6]

Barber, Brad M., Chip Heath, and Terrance Odean. 2003. "Good Reasons Sell: Reason-Based Choice Among Group and Individual Investors in the Stock Market." *Management Science* 49 (12): 1636–52. [5]

Bauer, Michal, Jana Cahlíková, Dagmara Celik Katreniak, Julie Chytilová, Lubomír Cingl, and Tomáš Želinský. 2024. "Nastiness in Groups." *Journal of the European Economic Association* 22 (5): 2075–107. [6]

Buffat, Justin, Matthias Praxmarer, and Matthias Sutter. 2023. "The Intrinsic Value of Decision Rights: A Replication and an Extension to Team Decision Making." *Journal of Economic Behavior & Organization* 209: 560–71. [5]

Charness, Gary, Ryan Oprea, and Sevgi Yuksel. 2021. "How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment." *Journal of the European Economic Association* 19 (3): 1656–91. [5]

Charness, Gary, and Matthias Sutter. 2012. "Groups Make Better Self-Interested Decisions." *Journal of Economic Perspectives* 26 (3): 157–76. [4, 6, 32]

Cooper, David J., and John H. Kagel. 2005. "Are Two Heads Better Than One? Team versus Individual Play in Signaling Games." *American Economic Review* 95 (3): 477–509. [5, 13, 21, 32]

DellaVigna, Stefano, and Devin Pope. 2018. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–56. [5]

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–29. [5]

Enke, Benjamin. 2020. "What You See Is All There Is." *The Quarterly Journal of Economics* 135 (3): 1363–98. [5]

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation Neglect in Belief Formation." *The Review of Economic Studies* 86 (1): 313–32. [3–5, 7, 9–11, 14, 16, 17, 31]

Fehrler, Sebastian, Anna Hochleitner, and Moritz Janas. 2026. "Too Good to Be True – Individual and Collective Decision-Making with Misleading Signals." *Management Science* 0 (0): null. [4, 6, 7]

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10 (2): 171–78. [13]

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–25. [13]

Haaland, Ingar, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2025. "Understanding Economic Behavior Using Open-Ended Survey Data." *Journal of Economic Literature* 63 (4): 1244–80. [25, 26, 45, 46]

Houghton, Susan M., Mark Simon, Karl Aquino, and Caren B. Goldberg. 2000. "No Safety in Numbers: Persistence of Biases and Their Effects on Team Risk Perception and Team Decision Making." *Group & Organization Management* 25 (4): 325–53. [6]

Keck, Steffen, Enrico Diecidue, and David V. Budescu. 2014. "Group Decisions under Ambiguity: Convergence to Neutrality." *Journal of Economic Behavior Organization* 103: 60–71. [5]

Kocher, Martin G., Matthias Praxmarer, and Matthias Sutter. 2020. "Team Decision-Making." In *Handbook of Labor, Human Resources and Population Economics,* edited by Klaus F. Zimmermann, 1–25. Cham: Springer. Living reference work entry. [4, 5, 15, 32]

Kocher, Martin G., Sabine Strauß, and Matthias Sutter. 2006. "Individual or Team Decision-Making—Causes and Consequences of Self-Selection." *Games and Economic Behavior* 56 (2): 259–70. [5]

Kocher, Martin G., and Matthias Sutter. 2005. "The Decision Maker Matters: Individual versus Group Behaviour in Experimental Beauty-Contest Games." *The Economic Journal* 115 (500): 200–223. [5]

Kugler, Tamar, Edgar Kausel, and Martin Kocher. 2012. "Are Groups More Rational than Individuals? A Review of Interactive Decision Making in Groups." *Wiley Interdiscip Review of Cognitive Science* 3 (4): 471–82. [4, 6, 32]

Lorge, Irving, and Herbert Solomon. 1955. "Two Models of Group Behavior in the Solution of Eureka-Type Problems." *Psychometrika* 20 (2): 139–48. [21]

Luhan, Wolfgang J., Martin G. Kocher, and Matthias Sutter. 2009. "Group Polarization in the Team Dictator Game Reconsidered." *Experimental Economics* 12 (1): 26–41. [5]

Maciejovsky, Boris, and David V. Budescu. 2007. "Collective Induction without Cooperation? Learning and Knowledge Transfer in Cooperative Groups and Competitive Auctions." *Journal of Personality and Social Psychology* 92 (5): 854–70. [5]

Mancuso, Vincent F., Victor S. Finomore, Katherine M. Rahill, Elizabeth A. Blair, and Gregory J. Funke. 2014. "Effects of Cognitive Biases on Distributed Team Decision Making." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58 (1): 405–9. [6]

Moser, Johannes, and Niklas Wallmeier. 2021. "Correlation Neglect in Voting Decisions: An Experiment." *Economics Letters* 198: 109656. [5]

Selten, Reinhard. 1998. "Axiomatic Characterization of the Quadratic Scoring Rule." *Experimental Economics* 1 (1): 43–62. [9]

Sutter, Matthias. 2007. "Are Teams Prone to Myopic Loss Aversion? An Experimental Study on Individual versus Team Investment Behavior." *Economics Letters* 97 (2): 128–32. [5]

Appendix A Supplementary Analysis

A.1 Additional Results

Table A.1. Difference-in-differences: absolute deviations from the rational benchmark

		TEAM			INDIVIDUAL				
		Abs. Deviation from Rational Benchmark		Diff	Abs. Deviation from Rational Benchmark		Diff		
True state	Rational benchmark	Uncorrelated	Correlated	C-U	Uncorrelated	Correlated	C-U	Diff-in-Diff	IND vs TEAM
10	7.75	1.25	1.50	0.25	0.75	0.85	0.10	-0.15	TEAM gap larger
88	71.25	3.75	9.75	6.00	1.95	11.75	9.80	3.80	IND gap larger
250	259.75	9.75	23.25	13.50	8.25	26.75	18.50	5.00	IND gap larger
732	853.25	33.25	59.25	26.00	33.25	95.75	62.50	36.50	IND gap larger
1000	974.75	50.25	55.34	5.09	25.25	67.45	42.20	37.12	IND gap larger
4698	4810.00	110.00	490.00	380.00	193.00	690.00	497.00	117.00	IND gap larger
7338	8604.50	1395.50	898.00	-497.50	1279.50	881.50	-398.00	-99.50	TEAM gap larger
10000	7232.25	192.75	982.25	789.50	249.75	1539.25	1289.50	500.00	IND gap larger
23112	26331.00	1690.90	4331.00	2640.10	1521.00	5335.25	3814.25	1174.15	IND gap larger
46422	38910.50	910.50	5139.00	4228.50	2065.50	8856.50	6791.00	2562.50	IND gap larger

Note: This table presents an overview of the absolute deviations of the stated posteriors from the rational benchmarks across the 10 estimation tasks. We report the median deviations within an estimation task for the TEAM and INDIVIDUAL treatments split by Uncorrelated and Correlated conditions. Using a Diff-in-Diff approach we first report the difference in deviations within the respective treatment but across conditions. In a second step, we compare these differences across treatments: Diff-in-Diff = |C-U| in INDIVIDUAL minus |C-U| in TEAM, so positive values indicate a larger gap between conditions for individuals (IND gap larger) and negative values a larger gap for teams (TEAM gap larger). All figures are descriptive medians without statistical inference. In 8 of 10 tasks the gap between conditions is larger for individuals, in 2 of 10 for teams.

Table A.2. Difference-in-differences: normalized absolute deviations from the rational benchmark

		TEAM			INDIVIDUAL				
		Abs. Deviation from Rational Benchmark		Diff	Abs. Deviation from Rational Benchmark		Diff		
True state	Rational benchmark	Uncorrelated	Correlated	C-U	Uncorrelated	Correlated	C-U	Diff-in-Diff	IND vs TEAM
10	7.75	0.161	0.194	0.032	0.097	0.110	0.013	-0.019	TEAM gap larger
88	71.25	0.053	0.137	0.084	0.027	0.165	0.138	0.053	IND gap larger
250	259.75	0.038	0.090	0.052	0.032	0.103	0.071	0.019	IND gap larger
732	853.25	0.039	0.069	0.030	0.039	0.112	0.073	0.043	IND gap larger
1000	974.75	0.052	0.057	0.005	0.026	0.069	0.043	0.038	IND gap larger
4698	4810.00	0.023	0.102	0.079	0.040	0.143	0.103	0.024	IND gap larger
7338	8604.50	0.162	0.104	-0.058	0.149	0.102	-0.046	-0.012	TEAM gap larger
10000	7232.25	0.027	0.136	0.109	0.035	0.213	0.178	0.069	IND gap larger
23112	26331.00	0.064	0.164	0.100	0.058	0.203	0.145	0.045	IND gap larger
46422	38910.50	0.023	0.132	0.109	0.053	0.228	0.175	0.066	IND gap larger

Note: This table presents an overview of the absolute deviations (normalized by the rational benchmark) of the stated posteriors from the rational benchmarks across the 10 estimation tasks. We report the median deviations within an estimation task for the TEAM and INDIVIDUAL treatments split by Uncorrelated and Correlated conditions. Using a Diff-in-Diff approach we first report the difference in deviations within the respective treatment but across conditions. In a second step, we compare these differences across treatments: Diff-in-Diff = |C-U| in INDIVIDUAL minus |C-U| in TEAM, so positive values indicate a larger gap between conditions for individuals (IND gap larger) and negative values a larger gap for teams (TEAM gap larger). All figures are descriptive medians without statistical inference. In 8 of 10 tasks the gap between conditions is larger for individuals, in 2 of 10 for teams.

Table A.3. Naivete by treatment and condition (band ±0.2)

	Uncorrelated		Correlated	
Statistic	IND	TEAM	IND	TEAM
Panel A: Descriptive statistics				
Mean	0.121	0.173	0.538	0.363
SD	0.398	0.261	0.530	0.420
Median	0.014	0.068	0.550	0.303
Share naive	0.013	0.043	0.225	0.159
Share rational	0.747	0.594	0.423	0.460
Panel B: p-values of tests				
	Kolmogorov-Smirnov	**Wilcoxon**	**Chi² (naive)**	**Chi² (rational)**
IND vs TEAM (Uncorrelated)	0.294	0.293	0.554	0.076
IND vs TEAM (Correlated)	0.130	0.061	0.451	0.791
Correlated vs Uncorrelated (INDIVIDUAL)	<0.001	<0.001	<0.001	<0.001
Correlated vs Uncorrelated (TEAM)	0.009	0.024	0.054	0.172

Note: This table reports and compares statistical moments of naivete parameters (subject-level) across treatments INDIVIDUAL and TEAM within the Uncorrelated and Correlated conditions. In the upper Panel A we report descriptive statistics and shares of subjects that align with fully rational/naive reporting behavior. We classify subjects as fully rational/naive if their median naivete (across tasks in Part 2) lies within 0.2 of 0 (rational) or 1 (naive). Panel B reports p-values of tests comparing the overall distribution (Kolmogorov-Smirnov), ranks (Wilcoxon rank-sum) and shares of fully rational/naive subjects (Chi²).

Table A.4. Pre-task (Part 1) classification, team composition and response times

	Team composition	N	Share (in %) rational in pre-task	Avg. response time (in min.)	Share (in %)
INDIVIDUAL					
Uncorrelated		75	72.6	1.8	100.0
Correlated		71	43.7	2.9	100.0
TEAM					
Uncorrelated	2 naive	3	0.0	3.5	4.5
Uncorrelated	1 naive, 1 rational	27	50.0	2.8	40.3
Uncorrelated	2 rational	37	100.0	3.2	55.2
Correlated	2 naive	20	0.0	3.2	32.3
Correlated	1 naive, 1 rational	35	50.0	4.3	56.5
Correlated	2 rational	7	100.0	4.5	11.3

Note: This table shows the share of subjects classified as rational in the solo pre-task (Part 1), i.e. with absolute pre-task naivete < 0.5. This classification determines team composition. The pre-task uses the information structure of the subject's condition, so the shares differ between conditions. Within each condition, the share of rational subjects does not differ between INDIVIDUAL and TEAM (72.6% vs. 75.7% in Uncorrelated and 43.7% vs. 39.2% in Correlated, Fisher's exact tests with p = 0.62 and 0.55). For teams, the shares follow from the composition by definition. N is the number of individuals or teams. Share is the share of individuals or teams within each treatment and condition. Three teams with an unclassified member are excluded. Average response times in minutes are computed per decision maker and task, excluding team-rounds without agreement.

Table A.5. OLS regressions: response time and chat length in teams

	Response time	**Chat length**	**Response time**
	(I)	(II)	(III)
Team Type 1	1.038***	0.576***	0.407
	(0.312)	(0.179)	(0.253)
Team Type 2	1.341***	0.454**	0.844**
	(0.394)	(0.187)	(0.346)
Chat length (log words)			1.095***
			(0.073)
Constant	3.003***	3.584***	-0.921**
	(0.627)	(0.352)	(0.420)
Num.Obs.	578	578	578
R2 Adj.	0.318	0.250	0.587
Task FE	Yes	Yes	Yes
Round FE	Yes	Yes	Yes
Session FE	Yes	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: OLS estimates with standard errors clustered at the team level in parentheses. TEAM, Correlated condition, one observation per team and task. Teams with an unclassified member and team-rounds without agreement are excluded. Reference category: teams with two naive members (Team Type 0). Response time in minutes. Chat length is the log of one plus the number of words in the team's chat in that task. All regressions exclude observations with absolute naivete > 3.

A.2 Alternative Definition of the Rational Answer

In the main text, a belief or a number in the chat counts as the rational answer if its implied naivete lies within ±0.2 of the rational benchmark ($|\chi| \leq 0.2$). Because this tolerance is a fixed share of the distance between the rational and the naive benchmark, it discriminates equally well in every task. Tables A.6 and A.7 repeat the analysis with a band of ±5% around the rational benchmark instead. This band is wide relative to the distance between the two benchmarks in tasks in which they lie close together (true states 1,000 and 7,338, where they differ by 7–8%) and narrow in tasks in which they lie far apart. The comparisons between team types are unchanged. Under the ±5% band, teams with one rational member state the rational belief significantly less often than rational individuals (−16.4 percentage points, $p < 0.05$). Under the main definition, the difference is smaller and not significant (−11.2 percentage points).

Table A.6. Rational beliefs of teams and simulated teams in the Correlated condition (rational belief within 5% of the benchmark)

		Simulated team (benchmark)		
	Team	**Averaging**	**Better member**	**Truth wins**
Type 0 (two naive)	12.2	12.9	18.0	32.5
Team minus benchmark (pp)		-0.7	-5.8	-20.3***
		(3.8)	(5.2)	(7.4)
Type 1 (one naive, one rational)	40.1	22.3	56.5	64.2
Team minus benchmark (pp)		+17.8***	-16.4**	-24.1***
		(5.9)	(7.5)	(7.0)
Type 2 (two rational)	65.7	45.4	56.5	81.0
Team minus benchmark (pp)		+20.2*	+9.2	-15.3
		(12.2)	(12.0)	(11.7)

Note: Correlated condition, one observation per decision maker (individual or team) and task. Share of tasks in which the decision maker states the rational belief (final belief within 5% of the rational benchmark), in percent. Rounds without agreement are excluded. Individuals are classified by their pre-task naivete (rational if absolute pre-task naivete < 0.5). Team type is the number of rational members. Simulated teams: within each task, all virtual pairs of individuals from the INDIVIDUAL treatment with the composition of the team type. The pair states the mean of the two beliefs (Averaging), the belief of the member with the better pre-task type (Better member, the rate of individuals of that type), or the rational belief if either member does (Truth wins). Rates average over pairs within a task, then over tasks. Cluster-bootstrap standard errors in parentheses (2000 replicates, with individuals resampled within type and teams within team type). * p < 0.1, ** p < 0.05, *** p < 0.01.

Table A.7. Generation and adoption of the rational answer in the team chat (rational answer within 5% of the benchmark)

	By team type			Difference between types	
	Type 0	**Type 1**	**Type 2**	**Type 1 − 0**	**Type 2 − 1**
Panel A: Correlated					
Generation	26.5	58.6	85.7	+32.1*** (6.2)	+27.1*** (7.0)
Adoption	45.3	68.6	77.2	+23.3** (10.1)	+8.6 (12.0)
Final belief rational	12.2	40.1	65.7	+27.9*** (6.0)	+25.6** (11.7)
Panel B: Uncorrelated					
Generation	63.3	73.0	76.2	+9.6 (23.2)	+3.3 (6.5)
Adoption	73.7	75.8	70.3	+2.1 (12.3)	-5.5 (7.8)
Final belief rational	50.0	55.3	53.7	+5.3 (14.6)	-1.5 (8.6)

Note: One observation per team and round. Teams are classified by the number of members who were rational in the pre-task (team type). The rational answer is a number within +/-5% of the rational benchmark. Generation: share of team-rounds in which the rational answer is mentioned in the chat (all numbers are extracted from the transcript, excluding numbers equal to a computer signal, an intermediary message under either information structure, or a sum of these). Adoption: share of these rounds in which the team's final belief is the rational answer. Final belief rational: share of all rounds in which the final belief is the rational answer. Adoption and final belief exclude rounds without agreement. Shares in percent by team type. Differences between adjacent team types in percentage points. Cluster-bootstrap standard errors in parentheses (teams resampled with replacement within condition and team type, 2000 replicates). *, ** and *** mark p < 0.1, 0.05 and 0.01 (normal approximation).

Appendix B Supplementary Materials: Coding the Team Chat Content

This appendix documents the coding procedure summarized in Section 4.1: the codebook induction and the resulting instrument, the human reference coding, the classification runs, the solution indicator, the chat-content comparisons and the prompts. Every reliability statistic is a deterministic function of the archived codings. The comparisons that resample — the condition and team-type comparisons of Appendix B.5, including the agreement between coders — use a cluster bootstrap over teams with 2,000 replicates and a fixed seed, in which a drawn team contributes all ten of its rounds.

B.1 Codebook Induction and the Consensus Instrument

Table B.1 gives the parameters of the three-stage induction. Table B.2 reproduces the consensus instrument as the coders and the models received it, with the source codebooks of each category. According to the merge model's provenance field, nine categories were proposed by two or more of the four model families and four by all four. In this sense, the instrument is a consensus rather than the output of one model.

Table B.1. Parameters of the codebook induction and merge procedure

Parameter	Value
Transcripts per induction run	50 (random draw from the 1,320)
Induction runs per model	10
Seeds	$42 + i$ for run i, giving a distinct subsample per run
Induction models	`mistral-large-3-675b-instruct-2512`, `gpt-oss-120b`, `llama-3.1-sauerkrautlm-70b-instruct`, `qwen3-30b-a3b-instruct-2507`
Merge model	`mistral-large-3-675b-instruct-2512`
Required fields per category	label (2–4 words), definition (2–3 sentences), representative quote, key distinguishing indicators
Reconciliation criterion	retain only categories that appear consistently across runs
Merge provenance fields	binary marker (1 = unique to a subset of codebooks, 0 = recurring pattern), source codebook IDs
Categories in the consensus codebook	15 (9 recurring across codebooks, 6 unique to a subset)

Notes: Stage 1 induces a codebook per model from ten independent random subsamples and reconciles the ten proposals into one codebook per model. Stage 2 merges the four model-specific codebooks into a single consensus codebook. Stage 3 hands the consensus codebook to the two human coders. The prompts of stages 1 and 2 and of the classification are reproduced in Appendix B.6.

Table B.2. The consensus codebook: 15 categories applied by human and machine coders

Label	Definition and illustrative quote	Source codebooks
Direct Calculation	The team explicitly computes the exact benchmark using arithmetic operations (e.g. summing values and dividing by count) without estimation, adjustment, or reliance on external anchors. *"Sum is 4000, divided by 4 is 1000."*	1, 2, 3, 4
Outlier Adjustment	The team identifies and discards or downweights extreme values before calculating the benchmark, often justifying the decision explicitly by deviation or perceived implausibility. *"The value 5000 is an outlier, let's leave it out."*	1, 2, 3, 4
Weighted Calculation	The team assigns different weights to values or applies a custom formula before computing the benchmark, often describing the weighting logic explicitly. *"We count the first value twice, then the average is 1200."*	1, 2, 4
Midpoint Compromise	The team selects a value between two proposed estimates without deeper calculation, justification, or iterative refinement. *"Let's take the midpoint between 800 and 1200, so 1000."*	1, 2, 4
Rounding & Simplification	The team rounds values or simplifies calculations to ease mental arithmetic, leading to approximate estimates and round numbers. *"The numbers are around 1000, let's take 1000."*	1, 2, 3, 4
Intuitive Guess	The team relies on gut feeling, rule of thumb, or rough estimation without explicit calculation, justification, or systematic method, often in vague or playful language. *"I'd guess somewhere around 8000."*	1, 2, 3, 4
Benchmark Anchoring	The team uses a provided benchmark, reference value, or external constraint as an anchor and adjusts its estimate relative to it, often without further manipulation. *"The benchmark is 5000, so we go a bit higher, 5500."*	1, 2, 4
Iterative Refinement	The team engages in multiple rounds of discussion, calculation, and correction, progressively refining its estimate through back-and-forth verification and recalculation. *"At first I thought 1200, but after the correction it's 1100."*	2, 4
Direct Agreement	The team quickly agrees on a solution without discussing or explaining its thought process, often on the basis of shared intuition, trust, or prior alignment. *"Yes, 1000 works."*	2, 3, 4
Humor & Banter	The dialogue contains jokes, laughing emoji, playful language, or sarcastic remarks unrelated to the numeric solution, often interspersed with task-related content. *"Haha, I can't do math, you take that one."*	2
Expressed Uncertainty	The team openly states doubt, asks for clarification, or admits lack of confidence in the calculation or method, often before proceeding to a solution. *"I'm not sure, why doesn't this work?"*	2
Procedural Coordination	The conversation focuses on task delegation, verification roles, or submission logistics rather than numeric content, often allocating computational steps. *"I'll compute the sum, you divide by 4."*	2
Social Small Talk	Remarks unrelated to the numeric solution, referring to the experimental setting, fatigue, time pressure, or personal anecdotes, without contributing to the task. *"This is the last round, I'm tired."*	2
Sum Verification	The team focuses on verifying the total sum of values before computing the benchmark, cross-checking the arithmetic explicitly. *"The sum is 4000, is that right?"*	1
Algorithmic Correction	The team applies a systematic rule or formula to adjust raw data before computing the benchmark, often with a step-by-step explanation. *"We double the value and subtract 200."*	1

Notes: The table reproduces the consensus codebook as it was given to the human coders and, verbatim, to the classifying models, except that the quotes are translated here from the German originals. Each entry also carries a list of key distinguishing indicators, omitted here for space and available in the replication material. The quotes are illustrative examples written by the merge model (stage 2) and make the classification prompt a few-shot prompt in the sense of Haaland et al. (2025). Source codebooks are the four model-specific codebooks of Table B.1. Categories appearing in a single source are those flagged as unique to a subset by the merge procedure.

B.2 Human Coding and Reference Standard

The rating sheet given to each coder contained, per transcript, the session, round and group identifiers, the round-specific rational benchmark with its lower and upper ±5% boundaries, the transcript itself, two binary accuracy judgments (*at least one proposal within the rational interval* and *all proposals outside the rational interval*), one column per category coded 1 for primary, 2 for secondary and 0 for absent, and a self-reported certainty on a 0–1 scale.

Both coders received the written codebook and were trained separately on the study setting, but not on the research question, and both were blind to the hypotheses throughout. The coders never interacted, and disagreements were resolved by a fixed rule rather than by discussion. This departs from the joint training and discussion-based reconciliation recommended by Haaland et al. (2025), who also note that where objectivity is the goal, few interactions between coders are preferable, because discussion can produce interpretative convergence.

The tie-break rule of Section 4.1 resolved 52 of the 132 transcripts by direct agreement on the primary category (tier A), 23 by one coder's primary matching the other's secondary (tier B), one by matching secondaries (tier C) and 56 by the certainty tiebreak (tier D).

B.3 Classification

Four open-weight models classified all 1,320 transcripts: Mistral Medium 3.5 (128B parameters), Gemma 4 (31B), Llama 3.3 (70B) and Qwen 3.6 (27B). They ran at temperature 0, which makes the classification deterministic and reproducible, and returned a structured JSON object with six fields: the primary category, a confidence rating, a one-sentence justification, an optional secondary category, and two yes/no answers on whether a proposal in the chat lies within, or all proposals lie outside, the rational interval (prompt in Appendix B.6).[10]

10. We do not use the two answers on the rational interval. Whether the rational answer is proposed in a chat is coded deterministically from the numbers in the transcript (Appendix B.4, see footnote 11).

B.4 The Solution Indicator

Whether the rational answer is proposed in a chat is a fact about the transcript and the benchmark, and we code it deterministically. We extract every number from the transcript (with a decimal comma or point, and with a point, comma or space as thousands separator) and discard numbers that equal one of the four computer signals or an intermediary's message under either information structure, or a sum of them. A round counts as proposing the rational answer if any remaining number has an implied naivete within ±0.2 of the rational benchmark ($|\chi| \leq 0.2$). Appendix A.2 uses a band of ±5% around the rational benchmark instead. The human coders judged the ±5% version on their rating sheets, so we validate the extraction with that band. On the 132 double-coded transcripts, the two coders agree with each other at $\kappa = 0.92$. The indicator agrees with coder A at $\kappa = 0.90$ and with coder B at $\kappa = 0.92$. On the 127 transcripts on which the coders agree, it matches their judgment in 97.6% of cases ($\kappa = 0.95$, sensitivity 1.00, specificity 0.94).

B.5 Chat Content across Conditions and Team Types

Table B.3 reports the comparison between conditions summarized in Section 4.2 in full: the gap in each category under each coder, a joint test of all categories, and each model's agreement with the reported coder. Under every model the category distribution differs significantly between conditions, but only modestly (Cramér's V between 0.15 and 0.19).

The four models cannot be ranked on the validation sample (all six paired exact McNemar tests give $p \geq 0.38$), so the bottom rows of the table ask whether the choice of coder matters for this comparison. No two coders find gaps of one percentage point or more with opposite signs, and the rank correlation between each model's gap vector and that of Qwen 3.6 ranges from 0.59 (Llama) to 0.91 (Gemma). The coders disagree most on two categories. For Direct Agreement, Gemma finds no gap where Qwen 3.6 finds −6.0 percentage points. For Humor & Banter, Mistral and Llama find gaps of −2.1 and −2.6 percentage points where Qwen 3.6 finds +0.2. Every model's largest difference from Qwen 3.6 falls in one of these two categories, with Holm-corrected p-values: Direct Agreement for Gemma ($p < 0.01$), and Humor & Banter for Mistral ($p = 0.036$) and

for Llama ($p = 0.079$). All other differences are smaller and not tested separately. This is why the main text interprets only category differences on which the coders agree.

Table B.4 repeats the team-type contrasts of Table 7 under each coder, and Figure B.1 shows the category shares by team type and by round.

Table B.3. Coded communication across conditions, by coder

	Share (%)		Corr minus Uncorr gap (pp)				
Category	Corr	Uncorr	Qwen	Mistral	Gemma	Llama	Average
Intuitive Guess	20.6	14.2	+6.4	+5.6	+4.1	+7.5	+5.9
Direct Agreement	21.0	27.0	-6.0	-3.8	+0.0	-8.4	-4.5
Direct Calculation	12.2	13.6	-1.4	+0.3	-2.3	+0.8	-0.6
Rounding & Simplification	7.0	5.5	+1.5	+0.3	+1.5	+2.5	+1.5
Outlier Adjustment	10.6	16.7	-6.0	-5.5	-5.5	-4.7**	-5.4
Midpoint Compromise	14.4	13.3	+1.1	+0.3	+0.4	-0.2	+0.4
Expressed Uncertainty	0.3	0.1	+0.2	+1.5	-0.1	+0.7	+0.6
Social Small Talk	0.3	0.1	+0.2	-0.3	+0.2	-0.1	-0.0
Weighted Calculation	0.3	0.7	-0.4	-0.4	-0.8	+0.3	-0.3
Procedural Coordination	1.1	0.3	+0.8	+1.6	+0.2	+0.3	+0.7
Humor & Banter	2.4	2.2	+0.2	-2.1	+0.0	-2.6	-1.1
Sum Verification	0.6	0.6	+0.1	-0.2	+0.0	-0.1	-0.1
Algorithmic Correction	1.7	0.0	+1.7	+1.1	+1.7	+0.0	+1.2
Benchmark Anchoring	0.3	0.3	+0.0	-0.4	+0.0	-0.2	-0.1
Iterative Refinement	7.0	5.4	+1.6	+2.0	+0.6	+4.1	+2.1
All categories jointly							
χ^2 test, p			<0.01	<0.01	<0.01	<0.01	
Cramér's V			0.18	0.17	0.15	0.19	
Agreement with Qwen							
Rank correlation ρ				0.78	0.91	0.59	
Sign flips				0	0	0	
Largest difference (pp)				-2.3**	+6.0***	-2.8*	

Notes: Shares: percent of transcripts in each category under Qwen 3.6 27B, the coder used in the analysis. Gaps: *Correlated* minus *Uncorrelated* under each model and their averaged prediction. Tests: cluster bootstrap over the 132 teams (2000 replicates), Holm-corrected across the 15 categories within each column. Gaps where one condition never shows the category are not tested. Joint test: Monte-Carlo χ^2 (5,000 replicates) with Cramér's V. Agreement: the Spearman correlation of the two gap vectors, the number of categories on which the two gaps differ in sign while both are at least one percentage point, and the largest difference between the two gaps (model minus Qwen), Holm-corrected across the three models. The largest differences concern Humor & Banter (Mistral, Llama) and Direct Agreement (Gemma). *, ** and *** mark $p < 0.1$, 0.05 and 0.01.

Table B.4. Coded communication by team type under each coder, Correlated condition

Category	Qwen 3.6	Mistral Medium 3.5	Gemma 4 31B	Llama 3.3 70B	Average
Type 1 vs 0					
Direct Agreement	-4.9	-4.4	-14.4	-11.6	-8.8
Intuitive Guess	-6.2	-9.0	+0.1	-1.6	-4.2
Midpoint Compromise	-2.6	+0.4	+1.2	+2.4	+0.3
Direct Calculation	+2.8	+4.1	+7.0	+3.6	+4.4
Outlier Adjustment	+6.4	+7.0	+5.6	+3.0	+5.5
Rounding & Simplification	+1.2	-0.5	-1.3	-0.9	-0.4
Iterative Refinement	-0.2	+0.6	-0.4	+3.6	+0.9
Humor & Banter	+1.4	+1.6	+0.3	+1.0	+1.1
Algorithmic Correction	+2.6	+1.4	+1.8	+0.0	+1.4**
Procedural Coordination	+0.9	+2.7	+0.3	-0.2	+0.9
Sum Verification	+0.6	+1.1	+0.0	+0.0	+0.4
Benchmark Anchoring	+0.6	-0.2	+0.6	-0.4	+0.1
Expressed Uncertainty	-0.5	-4.1	+0.0	-1.1	-1.4
Social Small Talk	-1.0	-0.9	-0.5	+0.1	-0.6
Weighted Calculation	-1.0	+0.1	-0.2	+2.3	+0.3
Type 2 vs 0					
Direct Agreement	-2.6	-4.9	-16.4	-9.9	-8.4
Intuitive Guess	-20.8**	-30.7***	-14.2**	-16.4	-20.5***
Midpoint Compromise	-14.6***	-9.6**	-5.1	-2.8	-8.0**
Direct Calculation	+10.5	+14.9	+25.3	+16.1	+16.7
Outlier Adjustment	+1.6	+3.0	+3.0	+1.3	+2.2
Rounding & Simplification	-2.2	-1.6	-6.4	-0.9	-2.8
Iterative Refinement	+21.2***	+17.5	+12.4	+13.0	+16.0***
Humor & Banter	+1.4	+2.8	+0.0	+0.7	+1.2
Algorithmic Correction	+2.9	+2.9	+2.4	+0.0	+2.0
Procedural Coordination	+0.9	+4.7	+0.0	-0.5	+1.3
Sum Verification	+2.9	+4.3	+0.0	+0.0	+1.8
Benchmark Anchoring	+0.0	-0.5	+0.0	-1.0	-0.4
Expressed Uncertainty	+0.9	-2.1	+0.0	-0.6	-0.4
Social Small Talk	-1.0	-0.1	-0.5	+0.9	-0.2
Weighted Calculation	-1.0	-0.5	-0.5	+0.0	-0.5
Type 2 vs 1					
Direct Agreement	+2.3	-0.6	-2.0	+1.7	+0.4
Intuitive Guess	-14.6**	-21.7***	-14.3***	-14.9	-16.4***
Midpoint Compromise	-12.0**	-10.0**	-6.3*	-5.1	-8.4**
Direct Calculation	+7.7	+10.9	+18.3	+12.6	+12.4
Outlier Adjustment	-4.9	-4.0	-2.6	-1.7	-3.3
Rounding & Simplification	-3.4	-1.1	-5.1	+0.0	-2.4
Iterative Refinement	+21.4***	+16.9	+12.9	+9.4	+15.1***
Humor & Banter	+0.0	+1.1	-0.3	-0.3	+0.1
Algorithmic Correction	+0.3	+1.4	+0.6	+0.0	+0.6
Procedural Coordination	+0.0	+2.0	-0.3	-0.3	+0.4
Sum Verification	+2.3	+3.1	+0.0	+0.0	+1.4
Benchmark Anchoring	-0.6	-0.3	-0.6	-0.6	-0.5
Expressed Uncertainty	+1.4	+2.0	+0.0	+0.6	+1.0
Social Small Talk	+0.0	+0.9	+0.0	+0.9	+0.4
Weighted Calculation	+0.0	-0.6	-0.3	-2.3	-0.8

Share under the higher team type minus share under the lower, in percentage points, Correlated condition, under each model and under their averaged prediction. Cluster bootstrap over teams, 2000 replicates. *, ** and *** mark $p < 0.1$, $p < 0.05$ and $p < 0.01$, each column tested on its own and Holm-corrected across the 15 categories within that column and contrast. The averaged prediction is tested by averaging the four models inside each bootstrap replicate. Differences where one of the two team types never shows the category are not tested, because the bootstrap distribution is then degenerate.

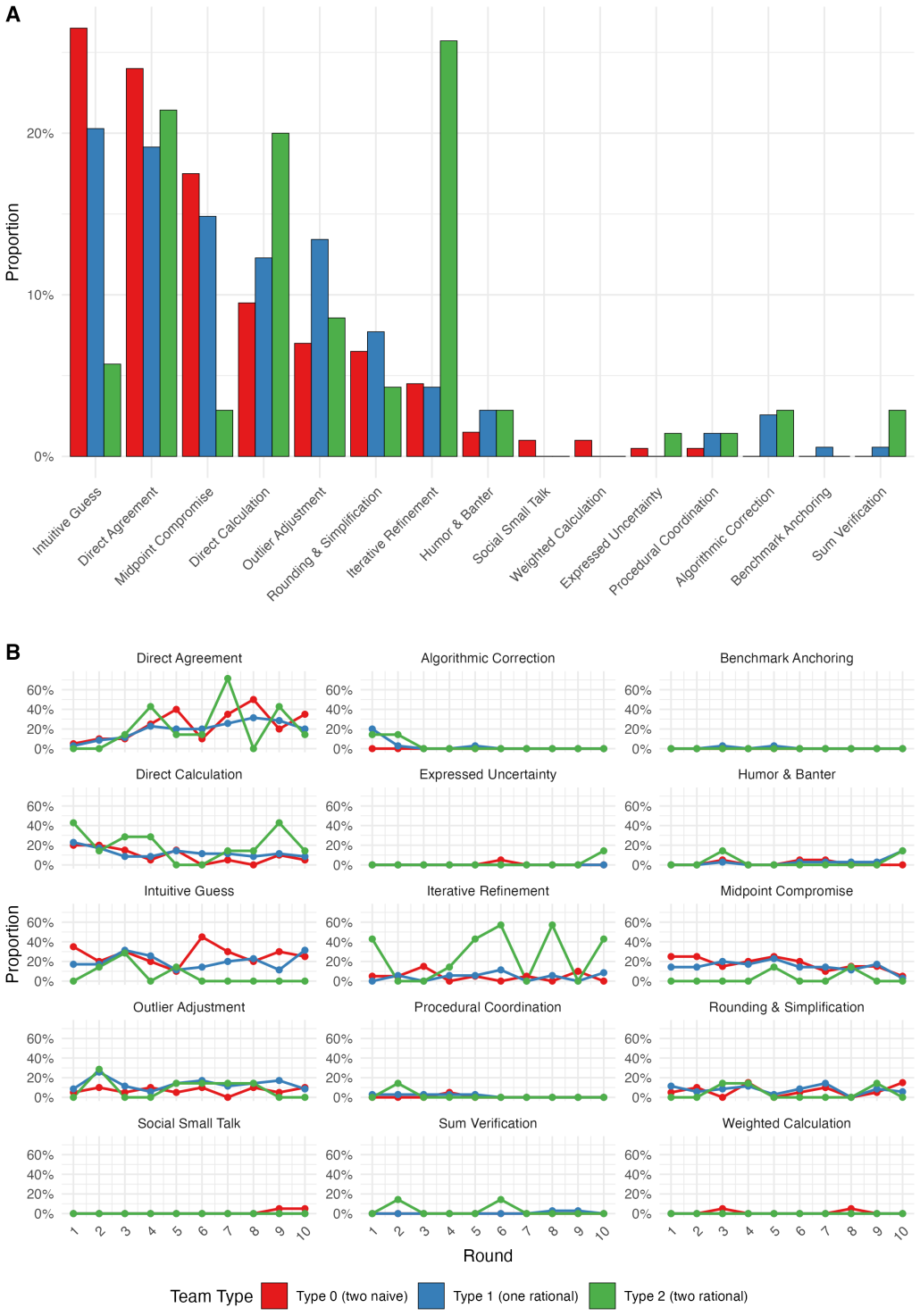


Figure B.1. Coded communication by team type in the *Correlated* condition: overall and by round

Notes: Panel A: share of each category by team type, pooled across rounds (as in Table 7). Panel B: share of each category by round and team type. Each Type 2 point rests on seven transcripts. Categories coded by Qwen 3.6.

B.6 Prompts

The four prompts are reproduced as they were sent. Placeholders in angle brackets are the model's fields to fill, except italic ones, which mark material inserted programmatically.

Stage 1: within-model induction. System message:

> *You are a research methodology expert.*

User message:

> *Analyze these chat transcripts and identify a set of mutually exclusive and exhaustive (MEE) categories describing the different solution strategies and communication patterns teams used. Aim for a reasonable amount of categories that are MEE.*
>
> *For each category provide ONLY:*
> *1. Short label (2-4 words)*
> *2. Precise definition (2-3 sentences max)*
> *3. One representative quote (short)*
> *4. Key distinguishing indicators*
>
> *TRANSCRIPTS:*
> [50 randomly drawn transcripts, each preceded by the line *[Transcript - Benchmark:* ⟨round-specific rational benchmark⟩ *]*]

Stage 1: reconciliation across the ten runs. System message:

> *You are a research methodology expert.*

User message:

> *You have run category induction on several different subsamples of the same dataset. Merge and reconcile the proposed coding schemes below into one final codebook. Retain only categories that appear consistently across runs. Resolve overlaps. Output a clean numbered codebook in the same format (label, definition, quote, key indicators). Don't output the reasoning process, any emojis or any documentation/description of the process!*
>
> [the ten run outputs, each preceded by *[Run* ⟨i⟩ *]*]

Stage 2: cross-model merge. The system message is that of the stage 1 induction. User message:

> *Analyze these codebooks for a text classification task. Identify a set of similar and unique categories. For categories to be similar they should describe the same*

kind of pattern across codebooks. For categories to be unique they should describe patterns that only appear in a subset of codebooks and are not captchured by any other of the given categories. Ultimately, provide one unifying codebook that resembles similar and unique patterns using the same structure as in the input codebooks.

That is, for each category provide ONLY:
1. Short label (2-4 words)
2. Precise definition (2-3 sentences max)
3. One representative quote (short)
4. Key distinguishing indicators
5. Binary marker (1=unique to subset of codebooks, 0=similar pattern across codebooks)
6. Codebook ID where the category can be found

Guidelines:
- Similar categories: describe same patterns across codebooks
- Unique categories: appear in only some codebooks, not captured elsewhere
- Maintain input codebooks' structure
- Prioritize clarity and specificity in definitions

[the four model-specific codebooks, each preceded by *[Codebook* $\langle i \rangle$ *]*]

Classification. System message, identical for all four models and all 1,320 transcripts:[11]

You are a research assistant coding experimental chat transcripts. Respond only with valid JSON, no other text. Don't output the reasoning process.
Apply the coding scheme below to the chat transcript, the Group ID and the round-specific rational benchmarks.
Assign exactly one primary category. Note a secondary category only if clearly applicable.

Complementary to the coding scheme, include these two additional categories about solution accuracy:1. Chat contains at least one suggestion within rational benchmark range: When any participant suggests a solution within ±5%% of the rational benchmark shown for each transcript2. Chat contains only suggestions outside benchmark range: When all suggested solutions are outside ±5%% of the rational benchmark shown for each transcriptCODING SCHEME:

[the consensus codebook of Table B.2, verbatim, as JSON]

*Respond **only** with the following JSON format, **no markdown, no code fences, no extra text, no reasoning**:*

```
{
  "primary_category": "<label>",
  "confidence": "high" | "medium" | "low",
```

11. The prompt also asks the models whether the chat contains a suggestion within ±5% of the rational benchmark. We elicited these answers but do not use them: whether a number in the transcript lies inside the rational band is a factual question, which the deterministic solution indicator of Appendix B.4 answers more precisely. On the 127 double-coded transcripts on which the two human coders agree, the indicator reaches Cohen's $\kappa = 0.95$ against their judgment (sensitivity 1.00, specificity 0.94), the four models $\kappa = 0.72$–0.81 (sensitivity 0.79–0.84).

```
  "justification": "<one sentence>",
  "secondary_category": "<label or null>",
  "solution_within": "<yes or no>",
  "solution_outside": "<yes or no>"
}
```

User message, one per transcript:

> *TRANSCRIPT:*
> [the transcript, messages concatenated with semicolons]
> *; Rational Benchmark:* ⟨benchmark⟩*; 5%-limit around rational benchmark: [*⟨0.95× benchmark⟩, ⟨1.05× benchmark⟩*]*

Appendix C Order of Belief Formation Tasks in Main Treatments

This is kept the same as EZ19 (Online Appendix) for the purposes of replication. All belief elicitation treatments are implemented in three different randomized orders of rounds. Ordered by true state, they are as follows:

(1) 10000, 88, 46422, 4698, 250, 23112, 1000, 10, 7338, 732

(2) 732, 23112, 88, 1000, 250, 4698, 10, 7338, 10000, 46422

(3) 250, 7338, 10000, 10, 4698, 88, 46422, 732, 1000, 23112

Appendix D Instructions

D.1 Treatment: TEAM – *Correlated*

Instructions (to be handed out to participants in hard copy). You will now take part in an economic experiment. You will receive an amount of 5 euros for your participation in the experiment. This will be paid to you at the end of the experiment. You can earn additional money, which is also paid out at the end of the experiment. How much you earn depends on your decisions. In this experiment we will talk about points. The points you collect during the experiment are exchanged for euros and paid out at the end of the experiment. Talking to other participants is not allowed during the experiment. If you have any questions, you can raise your arm outside the booth and the experimenter will attempt to answer your questions.

Your task. In this experiment you have to solve eleven guessing tasks. In **Part 1** you have to solve one **guessing task alone** and in **Part 2** you have to solve ten guessing tasks **together with a randomly selected partner in the team**. These tasks require you to estimate how many objects are in a given imaginary container. These objects can be, for example, peas in a vase or stones in a jar. All estimation tasks are different and completely independent of each other. Your earnings

depend on how accurately you estimate, that is, how close your estimate is to the actual number of objects in the container. At the end of the experiment, the guess from part 1 and a randomly chosen guess from part 2 will be payable and you will be paid based on the accuracy of your guesses in the tasks. This is explained in more detail on the following pages.

Your decision. See the final page for a screenshot of the decision screen. You can enter a number by typing the number in the blue box. **A decision is not valid until you have confirmed the estimate entered.**

Procedure for the individual estimation tasks. You are given a total of 7 minutes for each of the 11 estimation tasks. After the individual estimation, the start of the team part will be announced. In the team section, after 5 minutes you will receive a message on the screen that you only have 2 minutes to make your decision. For both team members, this means that a decision must be confirmed (click confirmation button) after a maximum of 7 minutes in order to have a valid estimate. However, you can also move on to the next estimation task at any time beforehand, provided that you have submitted a valid estimation.

Team decisions in part 2. For these estimation tasks, you are randomly paired with another participant (always the same person). You and your team partner can communicate with each other (anonymous team chat). **Both team members see the same screen and numbers (see last page) but have to enter estimates individually**. You can enter a number by typing the number in the blue field (your guess: the jar contained the following number of items) and clicking the "OK" button. You can change the number by entering a new number and clicking OK. At the end of each estimation task, your team must have made a decision (= both team partners have entered the same estimation individually). To finally confirm a number as a team decision, one of you must click the "**Confirm Estimation Task**" button. **A number cannot be confirmed until both team members have entered the same number with "OK."** If your team's entered estimates do not match, you will receive information on the screen.

Team chat. In each task, you and your partner can talk via **chat, which is only visible to your team**. The chat is displayed in the lower half of the screen for all tasks. You send a message by typing the text in the blue line and then confirming with the "Enter" key. The message is then immediately visible to you and your partner in the chat history. **The chat history is emptied after each guessing task.**

No threats or insults may be uttered in the chat and no information may be disclosed that allows conclusions to be drawn about your identity (e.g. name or seat number).

Your earnings. In addition to your show-up fee payment, you will be paid based on the accuracy of your estimates. In this experiment we will talk about points. 100 points correspond to 2 euros in part 1 (individual estimation) and 10 euros in part 2 (team estimations). The points you collect during the experiment are exchanged for euros and are also paid out at the end of the experiment. You get more money the closer your estimate is to the actual number of items in the bin. Your individual estimate and a randomly selected estimation task (from a total of 10 estimates) from part 2 (team estimates) are relevant for the payment. This means each team estimate is potentially relevant to your payment, so you should think carefully about each task. You can earn a maximum of 100 points per task. In the **team part, the 100 points are to be understood per person, which means that the payout for each team member is calculated according to the formula below**. You get these 100 points if you guess the number of objects correctly. The further your estimate is from the actual number, the less you will earn. This is determined using the following formula:

$$\textbf{Payment} = 100 - 0.16 \times \textbf{(percentage deviation of appraised value from real value)}^2$$

This means that the percentage deviation (in percent) between your estimate and the true value is squared and multiplied by 0.16. This number is then subtracted from the maximum earning of 100 points. Although this formula looks complicated, the underlying principle is very simple: the smaller the difference between your estimate and the true value, the larger your profit. However, your earnings can never be less than zero, i.e., you cannot lose money.

Consider the following example: Suppose said container contains 650 objects. You will then get the following number of points depending on your guess:

Estimate 0 → deviation 100% → earns 0 points
Estimate 520 → deviation 20% → earns 36 points
Estimate 585 → deviation 10% → earns 84 points
Estimate 650 → deviation 0% → earns 100 points
Estimate 715 → deviation 10% → earns 84 points
Estimate 780 → deviation 20% → earns 36 points
Estimate 1300 → deviation 100% → earns 0 points

You see that your earnings depend only on the percentage deviation. For example, it doesn't matter if you overestimate or underestimate the true contents of the container by 10 items.

Information regarding the estimation tasks. You will not see any of the estimation tasks, i.e. you won't see any of these imaginary containers. Instead, you receive various computer-generated information about the correct estimation result for each estimation task. For each task, you see this information and then enter your own estimate. The information you will receive is explained in detail below.

On computers, we simulate devices that solve exactly the same guessing tasks as you. There are two different types of devices. On the one hand there are devices that themselves give an estimate of the contents of the containers (these devices are called measuring devices and are identified by letters). On the other hand, there are devices that observe the estimates of the measuring devices and calculate their own estimate based on these reports (these devices are called communication devices and are identified by numbers). For each estimation task, 10,000 computers (measuring devices) made their own estimates; the estimates of these devices are completely independent of each other. The measuring devices all have the same quality, i.e., they are equally good at estimating. Note that these devices are good at solving these guessing tasks. So if these devices determined many estimates, then the average of those estimates would be correct.

So while almost every single guess contains an error, the average of those 10,000 guesses is correct (or very accurate). In addition, many estimates are relatively close to the true value. Below is an illustrative example in the summary.

Summary. (1) The measuring devices make mistakes, but the estimate is much more likely to be close to the true value than to be far from it.
(2) The average of these 10,000 estimates is correct. This means that the devices are correct (or very accurate) on average. Below we present an example.

Suppose the container contains 650 objects. The following figure illustrates how often the 10,000 measuring devices give which estimate. You can see that the frequency increases as you approach the center of the bell curve (the true value is 650). Therefore, it is much more likely that an estimate approaches the true value rather than deviating greatly from it. Also, you can deduce that the average of all estimates is 650, which is correct.

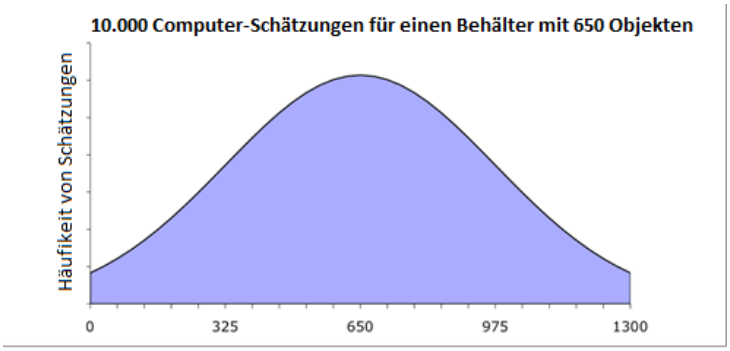


Figure 1: Sample computer estimates for a bin containing 650 items

For each estimation task, a computer randomly selects the estimates from exactly four measuring devices (A, B, C, D) from these 10,000 estimates. These four estimates are completely independent of each other. In addition to the measuring devices, there are three other communication devices (1, 2, 3). These communication devices do not determine their own estimation. Rather, they always observe two measuring devices and calculate an estimate from these estimates. More specifically, they do this by averaging these two estimates. You will get the following information as described in Figure 2.

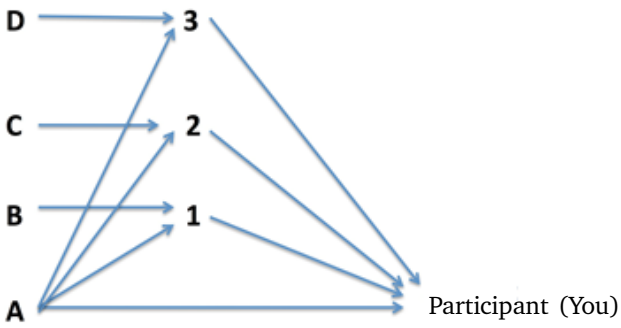


Figure 2: The four measuring devices determine their own estimates. The three communication devices receive the estimate of measuring device A and another measuring device as described by the arrows. They calculate their own estimate by averaging the two estimates. You can see the estimate of measuring device A as well as the three communication devices.

This means that you get the following information: You get the estimate of measuring device A as well as the estimates of all communication devices. As shown in Figure 2, each communication device sees the estimates of measuring device A and another measuring device. As you can see, communicator 1 receives the estimates from measuring devices A and B. Communicator 2 sees the estimates from measuring devices A and C. Communicator 3 sees the estimates from measuring devices A and D. As discussed earlier, the communicators average these two estimates and report that average as their estimate.

An example. You get the estimate of measuring device A, as well as the estimates of communicators 1, 2, and 3. The following simple example illustrates this. Above, we showed you the 10,000 estimates for the 650-object container. Now imagine that from these estimates, a computer randomly selects the following estimates from the estimators:

Measuring device A: 810
Measuring device B: 126
Measuring device C: 1078
Measuring device D: 937

As described above, the communication devices take the average of two of these estimates. You then get the following estimates:

Communication device 1: 468
Communication device 2: 944
Communication device 3: 873.5

So for this guessing task, you would see the following information on your computer screen (see Figure 3):



Figure 3: Example screenshot

IMPORTANT! Confirm chat entries with the **Enter key on the keyboard.**

Both team members see the same numbers.

Each round, they enter estimates individually by clicking the "OK" button. A new number can be entered by clicking "OK" again.

If you and your team partner agree, confirm the number with the "confirmation button." Confirmation is only possible if the individual estimates are identical.

Please read these instructions carefully again. The experimenter will shortly read a summary of the instructions. You will then answer a series of comprehension questions on the computer to ensure that you have understood the instructions.

D.2 Treatment: TEAM – *Uncorrelated*

Notes: The instructions for TEAM Uncorrelated proceed exactly as for TEAM Correlated (see above Appendix D.1) with the exception of the following section:

For each estimation task, a computer randomly selects the estimates from exactly four measuring devices (A, B, C, D) from these 10,000 estimates. These four estimates are completely independent of each other. In addition to the measuring devices, there are three other communication devices (1, 2, 3). These communication devices do not determine their own estimation. Rather, they are always watching a measuring device and reporting that device's estimate. You will get the following information as described in Figure 2.

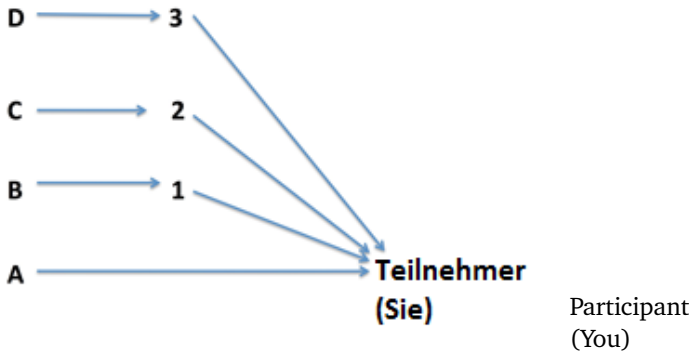


Figure 2: The four measuring devices determine their own estimates. The three communication devices receive a measuring device's estimate as described by the arrows. They calculate their own estimate by taking the measuring device's estimate. You can see the estimate of measuring device A as well as the three communication devices 1, 2, and 3.

This means that you get the following information: You get the estimate of measuring device A as well as the estimates of all communication devices. As shown in Figure 2, each communication device sees a measuring device's estimate. As you can see, communicator 1 receives measuring device B's estimate. Communicator 2 sees measuring device C's estimate. Communicator 3 sees measuring device D's estimate. As previously discussed, each communicator reports the estimate of one measuring device.

An example. You get the estimate of measuring device A, as well as the estimates of communicators 1, 2, and 3. The following simple example illustrates this. Above, we showed you the 10,000 estimates for the 650-object container. Now imagine that from these estimates, a computer randomly selects the following estimates from the estimators:

Measuring device A: 810
Measuring device B: 126

Measuring device C: 1078
Measuring device D: 937

As described above, each communication device reports the estimate of a measuring device. So they give the following estimates:

Communication device 1: 126
Communication device 2: 1078
Communication device 3: 937

So for this guessing task, you would see the following information on your computer screen (see Figure 3):

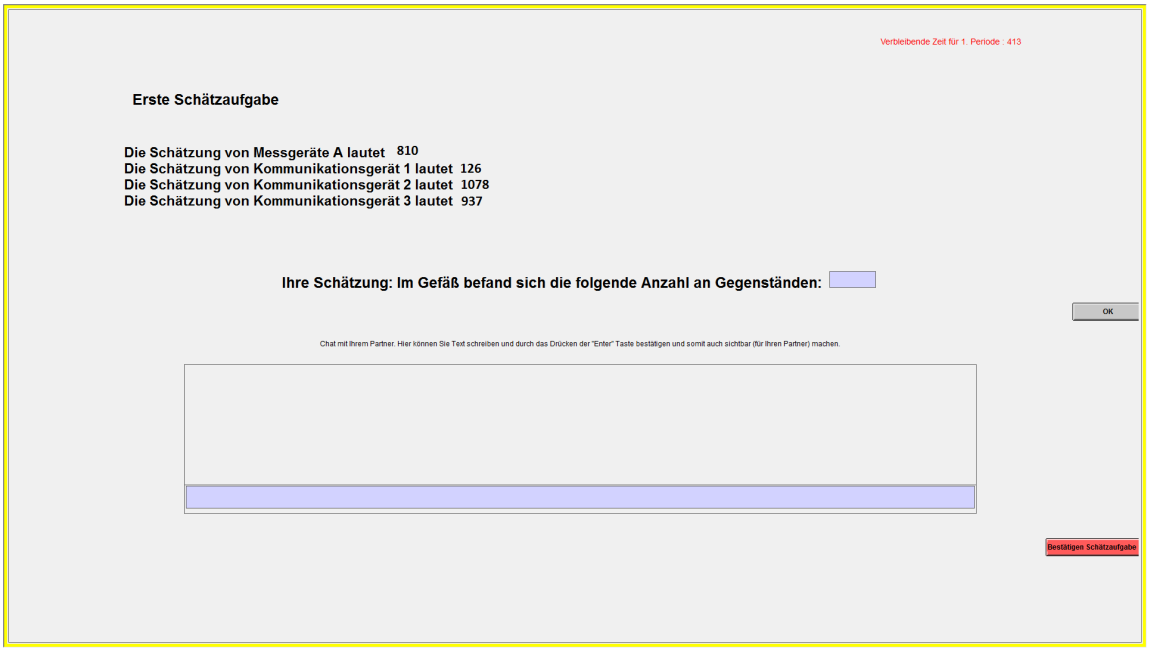


Figure 3: Example screenshot

D.3 Summary of Instructions for TEAM Treatments Only

Summary (to be read out loud by an experimenter).

- In this experiment you will be presented with 11 different estimation tasks.
- Part 1 is an individual decision and Part 2 consists of 10 team decisions.

- Both team members see the same numbers in the team decisions.
- At the end of the experiment, the task from Part 1 and one of the 10 rounds from Part 2 will be randomly selected to determine your earnings. Your earnings per task depend on how precise your answer was in the guessing task, that is, how close it was to the right value. Since all estimation tasks are potentially payout-relevant, you should answer carefully in all rounds.
- You won't actually see the guessing tasks. Instead, you receive computer-generated information about the solution to the estimation tasks.
- The structure of the information is as follows: In each round, four measuring devices that try to solve your guessing task are randomly selected. These devices are always equally good at solving the task. Each device provides its own estimate.
- In addition, there are 3 communication devices that process the estimates they observe from the estimators and then return an estimate derived from those estimates.
- Please look again at Figure 2 in your instructions. There you can see which estimates each communication device is observing, how they are processed, and how they are incorporated into the estimates reported by communication devices.
- For each estimation task you get the following information: You see the estimation of the measuring device A, as well as the estimations of communication devices 1, 2 and 3.
- After observing this information, you have 7 minutes to think about your own estimate and enter it into your computer.

D.4 Treatment: INDIVIDUAL – *Correlated*

Instructions (to be handed out to participants). You will now take part in an economic experiment. You will receive an amount of 5 euros for your participation in the experiment. This will be paid to you at the end of the experiment. You can earn additional money, which is also paid out at the end of the experiment. How much you earn depends on your decisions. In this experiment we will talk about points. The points you collect during the experiment are exchanged for euros and paid out

at the end of the experiment. Talking to other participants is not allowed during the experiment. If you have any questions, you can raise your arm outside the booth and the experimenter will attempt to answer your questions.

Your task. In this experiment you have to solve eleven guessing tasks. In **part 1** you have to solve one guessing task and then in **part 2** ten guessing tasks. These tasks require you to estimate how many objects are in a given imaginary container. These objects can be, for example, peas in a vase or stones in a jar. All estimation tasks are different and completely independent of each other. Your earnings depend on how accurately you estimate, i.e. how close your estimate is to the actual number of objects in the container. At the end of the experiment, the guess from part 1 and a randomly chosen guess from part 2 will be payable and you will be paid based on the accuracy of your guesses in the tasks. This is explained in more detail on the next page.

Your decision. See the final page for a screenshot of the decision screen. You can enter a number by typing the number in the blue box and clicking the "Next" button. Only click on "Next" if you really want to send the estimate, since no changes can be made afterwards. **A decision is only valid once you have confirmed the entered estimate with "Next."**

Procedure for the individual estimation tasks. You are given a total of **7 minutes** for each estimation task, and after 5 minutes you will receive information about the remaining time in the upper right part of the screen. For you, this means that a decision must be confirmed after a maximum of 7 minutes in order to have a valid (payout-relevant) estimate. However, you can also move on to the next estimation task at any time beforehand, provided that you have submitted a valid estimation.

Your earnings. In addition to your show-up fee earnings, you will be paid based on the accuracy of your estimates. In this experiment we will talk about points. 100 points correspond to 2 euros in part 1 and 10 euros in part 2. The points that you collect in the course of the experiment are exchanged for euros and are also paid out at the end of the experiment. You get more money the closer your estimate is to the actual number of items in the bin. Your estimate from Part 1

and a randomly selected estimation task (from a total of 10 estimates) from Part 2 are relevant for the payout. This means each estimate is potentially relevant to your payment, so you should think carefully about each task. You can earn a maximum of 100 points with one task. You get these 100 points if you guess the number of objects correctly. The further your estimate is from the actual number, the less you will earn.

This is determined using the following formula (per task):

$$\textbf{Payment} = 100 - 0.16 \times \textbf{(percentage deviation of estimated value from real value)}^2$$

This means that the percentage deviation (in percent) between your estimate and the true value is squared and multiplied by 0.16. This number is then subtracted from the maximum earning of 100 points. Although this formula looks complicated, the underlying principle is very simple: the smaller the difference between your estimate and the true value, the larger your profit. However, your earnings can never be less than zero, i.e., you cannot lose money.

Consider the following example: Suppose said container contains 650 objects. You will then get the following number of points depending on your guess:

Estimate 0 → deviation 100% → earns 0 points
Estimate 520 → deviation 20% → earns 36 points
Estimate 585 → deviation 10% → earns 84 points
Estimate 650 → deviation 0% → earns 100 points
Estimate 715 → deviation 10% → earns 84 points
Estimate 780 → deviation 20% → earns 36 points
Estimate 1300 → deviation 100% → earns 0 points

You see that your earnings depend only on the percentage deviation. For example, it doesn't matter if you overestimate or underestimate the true contents of the container by 10 items.

Information regarding the estimation tasks. You will not see any of the estimation tasks, i.e. you won't see any of these imaginary containers. Instead, you receive various computer-generated information about the correct estimation result for each estimation task. For each task, you see this information and then enter your own estimate. The information you will receive is explained in detail below.

On computers, we simulate devices that solve exactly the same guessing tasks as you. There are two different types of devices. On the one hand there are devices that themselves give an estimate of the contents of the containers (these devices are called measuring devices and are identified by letters). On the other hand, there are devices that observe the estimates of the measuring devices and calculate their own estimate based on these reports (these devices are called communication devices and are identified by numbers). For each estimation task, 10,000 computers (measuring devices) made their own estimates; the estimates of these devices are completely independent of each other. The measuring devices all have the same quality, i.e., they are equally good at estimating. Note that these measuring devices are good at solving these guessing tasks. So if these devices determined many estimates, then the average of those estimates would be correct.

So while almost every single guess contains an error, the average of those 10,000 guesses is correct (or very accurate). In addition, many estimates are relatively close to the true value. Below is an illustrative example in the summary.

Summary. (1) The measuring devices make mistakes, but the estimate is much more likely to be close to the true value than to be far from it.
(2) The average of these 10,000 estimates is correct. This means that the devices are correct (or very accurate) on average. Below we present an example.

Suppose the container contains 650 objects. The following figure illustrates how often the 10,000 measuring devices give which estimate. You can see that the frequency increases as you approach the center of the bell curve (the true value is 650). Therefore, it is much more likely that an estimate approaches the true value rather than deviating greatly from it. Also, you can deduce that the average of all estimates is 650, which is correct.

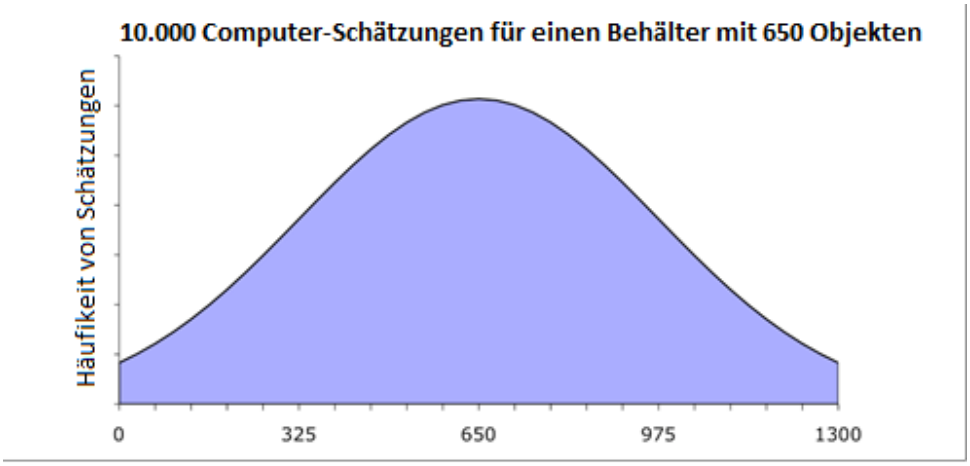


Figure 1: Sample computer estimates for a bin containing 650 items

For each estimation task, a computer randomly selects the estimates from exactly four measuring devices (A, B, C, D) from these 10,000 estimates. These four estimates are completely independent of each other. In addition to the measuring devices, there are three other communication devices (1, 2, 3). These communication devices do not determine their own estimation. Rather, they always observe two measuring devices and calculate an estimate from these estimates. More specifically, they do this by averaging these two estimates. You will get the following information as described in Figure 2.

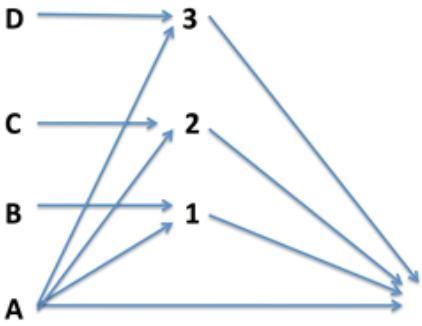


Figure 2: The four measuring devices determine their own estimates. The three communication devices receive the estimate of measuring device A and another measuring device as described by the arrows. They calculate their own estimate by averaging the two estimates. You can see the estimate of measuring device A as well as the three communication devices.

This means that you get the following information: You get the estimate of measuring device A as well as the estimates of all communication devices. As shown in Figure 2, each communication device sees the estimates of measuring device A and another measuring device. As you can see, communicator 1 receives the estimates from measuring devices A and B. Communicator 2 sees the estimates from measuring devices A and C. Communicator 3 sees the estimates from measuring devices A and D. As discussed earlier, the communicators average these two estimates and report that average as their estimate.

An example. You get the estimate of measuring device A, as well as the estimates of communicators 1, 2, and 3. The following simple example illustrates this. Above, we showed you the 10,000 estimates for the 650-object container. Now imagine that from these estimates, a computer randomly selects the following estimates from the estimators:

Measuring device A: 810
Measuring device B: 126
Measuring device C: 1078
Measuring device D: 937

As described above, the communication devices take the average of two of these estimates. You then get the following estimates:

Communication device 1: 468
Communication device 2: 944
Communication device 3: 873.5

So for this guessing task, you would see the following information on your computer screen (see Figure 3):



Figure 3: Example screenshot

Please read these instructions carefully again. The experimenter will shortly read a summary of the instructions. You will then answer a series of comprehension questions on the computer to ensure that you have understood the instructions.

D.5 Treatment: INDIVIDUAL – *Uncorrelated*

Notes: The instructions for INDIVIDUAL Uncorrelated proceed exactly as for INDIVIDUAL Correlated (see above Appendix D.4) except for the following section:

For each estimation task, a computer randomly selects the estimates from exactly four measuring devices (A, B, C, D) from these 10,000 estimates. These four estimates are completely independent of each other. In addition to the measuring devices, there are three other communication devices (1, 2, 3). These communication devices do not determine their own estimation. Rather, they are always watching a measuring device and reporting that device's estimate. You will get the following information as described in Figure 2.

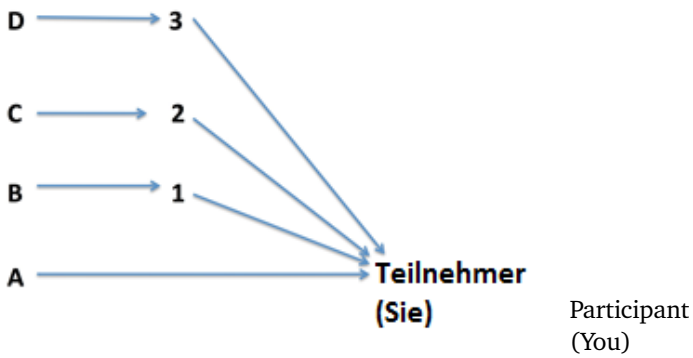


Figure 2: The four measuring devices determine their own estimates. The three communication devices receive a measuring device's estimate as described by the arrows. They calculate their own estimate by taking the measuring device's estimate. You can see the estimate of measuring device A as well as the three communication devices 1, 2, and 3.

This means that you get the following information: You get the estimate of measuring device A as well as the estimates of all communication devices. As shown in Figure 2, each communication device sees a measuring device's estimate. As you can see, communicator 1 receives measuring device B's estimate. Communicator 2 sees measuring device C's estimate. Communicator 3 sees measuring device D's estimate. As previously discussed, each communicator reports the estimate of one measuring device.

An example. You get the estimate of measuring device A, as well as the estimates of communicators 1, 2, and 3. The following simple example illustrates this. Above, we showed you the 10,000 estimates for the 650-object container. Now imagine that from these estimates, a computer randomly selects the following estimates from the estimators:

Measuring device A: 810
Measuring device B: 126
Measuring device C: 1078
Measuring device D: 937

As described above, each communication device reports the estimate of a measuring device. So they give the following estimates:

Communication device 1: 126
Communication device 2: 1078
Communication device 3: 937

So for this guessing task, you would see the following information on your computer screen (see Figure 3):

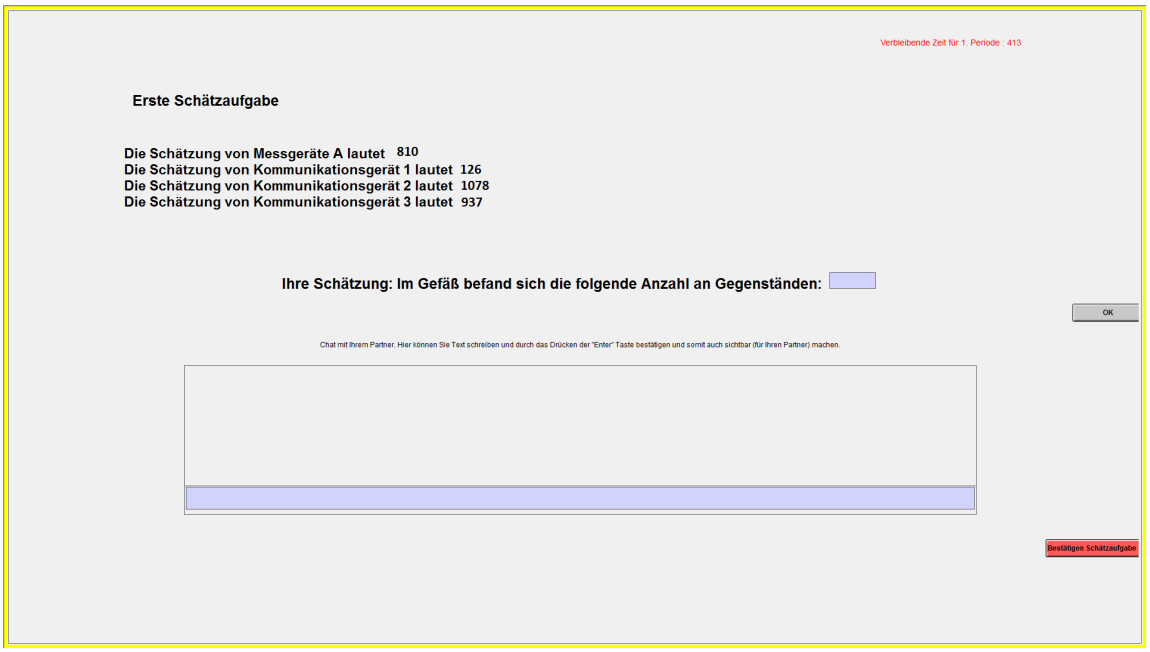


Figure 3: Example screenshot

D.6 Summary of Instructions for INDIVIDUAL Treatments Only

Summary (to be read out loud by an experimenter).

- In this experiment you will be confronted with 11 different estimation tasks.
- At the end of the experiment, the task from Part 1 and one of the 10 rounds from Part 2 will be randomly selected to determine your earnings. Your earnings per task depend on how accurate your answer was in the guessing task, that is, how close it was to the right value. Since all estimation tasks are potentially payout-relevant, you should answer carefully in all rounds.
- You won't actually see the guessing tasks. Instead, you receive computer-generated information about the solution to the estimation tasks.

- The structure of the information is as follows: In each round, four measuring devices that try to solve your guessing task are randomly selected. These devices are always equally good at solving the task. Each device provides its own estimate.
- In addition, there are three communication devices that process the estimates they observe from the estimators and then return an estimate derived from those estimates.
- Please look again at Figure 2 in your instructions. There you can see which estimates each communication device is observing, how they are processed, and how they are incorporated into the estimates reported by communication devices.
- For each estimation task you get the following information: You see the estimation of measuring device A as well as the estimations of communication devices 1, 2 and 3.
- After observing this information, you have 7 minutes to think about your own estimate and enter it into your computer.