Consensus and Truth-Finding in Narrative Communication*

Simin He† Sherry Xin Li‡ Junze Sun§ Xinlu Zou¶

June 10, 2026

Abstract

We study whether communication helps people reach agreement and correctly identify the narrative that generated observed evidence. In our laboratory experiment, one of two narratives about the gender wage gap is randomly selected as the true data-generating process. Participants observe independently drawn private data from that process and choose which narrative they believe is true before and after communicating anonymously with a matched partner who faces the same true narrative. We compare free-form communication with direct data sharing and a pre-communication reasoning task. Direct data sharing removes friction in transmitting private evidence, and pre-communication reasoning prompts participants to articulate their inference strategy before discussion. We find that free-form communication generates high consensus among initially disagreeing pairs but little improvement in true-narrative correctness. Data sharing shifts choices toward pair-level evidence, whereas reasoning improves selection of the narrative best supported by realized data but reduces consensus. The findings show that agreement and truth-finding can diverge in narrative communication.

Keywords: Narratives, Communication, Belief updating, Experiment.

JEL codes: C91, D83, D91.

*We thank the audiences at the 2025 ESA North America Meeting (Tucson), 2025 ESA World Meeting (Beijing), Duke Kunshan University, University of Amsterdam, NYU Shanghai, Sun Yat-Sen University, and Lingnan University for comments and feedback. Simin He acknowledges the National Natural Science Foundation of China (No. 72473089). Junze Sun acknowledges the National Natural Science Foundation of China (No. 72403185 and No. 72433003). The experiment is preregistered (AEARCTR-0016041). All errors are our own.

†School of Economics, Shanghai University of Finance and Economics, 777 Guoding Rd, 200433 Shanghai, China. E-mail: he.simin@mail.shufe.edu.cn

‡Department of Economics, University of Arkansas, 1 University of Arkansas, Fayetteville, AR 72701, United States. E-mail: SLi@walton.uark.edu

§School of Economics and Management, Wuhan University, 299 Bayi Road, 430072 Wuhan, China. E-mail: sunjunze@gmail.com

¶School of Finance, Dongbei University of Finance and Economics, 217 Jianshan Street, 116025 Dalian, Liaoning, China. E-mail: xinlu_zou@dufe.edu.cn

1 Introduction

People use narratives to interpret real-world evidence, and disagreement over factual issues often arises when they disagree about which narrative best explains the observed evidence. In debates about the causes of economic inequality, the consequences of public policies, or the risks of new technologies, people may observe similar facts yet draw different conclusions. Such disagreement is not merely a matter of taste or ideology. In many cases, the underlying issue has an objective answer, but people rely on different narratives to organize and interpret the available evidence. These competing interpretations can obstruct collective decision-making, reduce the effectiveness of policy communication, and sustain or exacerbate social and political polarization.

Communication is often proposed as a natural remedy. By talking to one another, people may exchange information, correct sampling errors, and learn how others reason from data. Yet ordinary communication may also fall short. Individuals often form narratives from limited private observations. They see different pieces of evidence, draw on different examples, and rely on different mental models to interpret what they observe. These observations are difficult to transmit fully, and the narratives that organize them are often implicit, incomplete, or hard to articulate.

This paper studies whether communication helps people reach agreement and identify the truth in a controlled laboratory experiment. We focus on communication about narratives, which are interpretive frameworks that organize facts and provide an account of how the world works. Narratives may be transmitted through simple stories, examples, metaphors, or social communication (Shiller, 2017). In many settings, multiple narratives can coexist without an unambiguous ground truth. In the recent economic literature, however, narratives are often modeled as mental models or data-generating processes, so that one narrative can be objectively more accurate than another (Eliaz and Spiegler, 2020). We study a setting in which participants may initially disagree about which narrative is correct, but there is an unknown true underlying data-generating process shared by all participants.

Communication can help in this setting through two channels. First, it can transmit private observations. When individuals observe independent private samples, disagreement may arise simply because their samples differ. The combination of observations could reduce sampling error and help participants identify the true underlying

narrative. Second, communication can transmit ways of thinking. Even when people observe similar data, they may disagree because they use different mental models to interpret those data. Explaining one's reasoning or understanding a partner's reasoning may therefore improve the quality of inference. These two channels are distinct, and they face different frictions. On the one hand, data are high-dimensional and may be costly to describe in ordinary conversation. On the other hand, mental models are abstract and may be difficult to convey clearly. Our design separates these channels by comparing unstructured communication with two structured modifications, one that removes the friction in data sharing, and the other that encourages participants to reflect on and better articulate their reasoning before communicating.

We implement this design in a laboratory experiment on narrative formation about the gender wage gap. Participants observe data generated by one of two competing narratives. Narrative 1 states that most of the observed gender wage gap cannot be explained by observable productivity differences, whereas narrative 2 states that roughly half of the gap is explained by such differences. Both narratives are based on published empirical studies and are presented as possible data-generating processes. At the beginning of each session, one narrative is randomly selected as the true narrative for each matching group in that session. All participants in the same matching group face the same true narrative, but each receives an independently drawn personal data graph. Participants first make an individual narrative choice in Stage 1, then communicate in randomly matched pairs before making a second narrative choice in Stage 2. Choices in both stages are incentivized by whether they identify the true narrative.

The experiment has three between-subject treatments. In *Communication*, participants chat freely for up to fifteen minutes while observing their own data graph and both participants' Stage 1 choices. In *Share*, participants additionally observe their partner's data graph and the pair's combined data graph, which removes the need to transmit private data verbally. In *Reasoning*, participants first write incentivized advice for a future participant facing the same inference problem, and then enter the same chat environment as in *Communication*. This treatment encourages participants to articulate their reasoning before communication. Its effect is ambiguous ex ante, since better articulation may facilitate learning but may also increase participants' attachment to their initial interpretation.

The experiment yields three main findings. First, communication produces high consensus but limited improvement in narrative choice accuracy. Among pairs who initially choose different narratives, more than 80% reach consensus in *Communication*. This high consensus rate, however, is not accompanied by a significant increase in true-narrative correctness, nor does it reliably improve best-fit correctness with the available data.[1] Consensus is therefore distinct from truth-finding. Communication can coordinate participants on a common narrative without ensuring that the agreed-upon narrative is the correct or best supported by evidence. This finding is particularly striking because our experiment incentivizes truth-finding, not consensus.

Second, reducing data-sharing frictions improves how participants use group evidence, but does not fully solve the truth-finding problem. In *Share*, consensus remains high, and we find evidence that participants are more likely to choose the narrative that best fits the pair's combined data. At the same time, true-narrative correctness does not improve relative to *Communication*. This may reflect sampling noise since the narrative that best fits the realized data may not coincide with the true data-generating process. Thus, best-fit correctness captures evidence-based inference, whereas true-narrative correctness captures recovery of the underlying truth. Direct data sharing therefore helps participants align their choices with observed group evidence, but does not mechanically ensure convergence to the underlying truth.

Third, prompting participants to articulate their reasoning improves evidence-based correctness but reduces consensus. In *Reasoning*, consensus among initially conflicting pairs falls to 52%, well below the rates in *Communication* and *Share*. Yet participants are more likely to choose the narrative that best fits their private data. Moreover, conditional on reaching consensus, conflicting pairs are more likely to converge on the narrative with stronger empirical fit. The mechanism analyses support this interpretation, indicating that participants in *Reasoning* are more likely to stand by their private evidence and more often describe their data and reasoning in chat. Thus, *Reasoning* appears to discipline evidence evaluation while making narratives harder to revise.

In summary, the results reveal an inherent tension in narrative communication. Ideally, communication would both generate agreement and move participants closer to

[1] The true narrative is the underlying narrative selected by the computer to generate the data graph. The best-fit narrative is the narrative best supported by the realized observed data.

the truth. We find that free-form communication achieves the first goal but not the second. Direct data sharing helps participants incorporate combined evidence, while explicit reasoning improves selection of better-fitting narratives. However, reasoning also reduces consensus, suggesting that articulated mental models may be harder to revise without sufficiently strong counter-evidence or argument. Thus, these results show that different forms of communication operate on distinct margins, with data sharing improving access to combined evidence and reasoning enhancing interpretation, even though neither improvement necessarily translates into higher true-narrative correctness.

This paper contributes to the growing literature on narratives, mental models, and communication. A large literature studies how narratives shape beliefs and behavior (Shiller, 2017; Eliaz and Spiegler, 2020), while recent work examines how narratives can be strategically used or tailored to persuade audiences (Schwartzstein and Sunderam, 2021; Aina, 2023; Barron and Fries, 2025). Our contribution is to study non-strategic communication between participants who share a common objective but may differ in private observations and interpretive models. The laboratory environment allows us to observe both consensus and correctness relative to an objective truth, and to distinguish true-narrative correctness from best-fit correctness based on realized data. This distinction is crucial to our study since we demonstrate that communication can generate agreement without improving truth-finding, while interventions that strengthen evidence-based reasoning may reduce agreement.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the experimental design and hypotheses. Section 4 reports and discusses the experimental results. Section 5 concludes.

2 Related Literature

This paper connects several strands of literature on narratives, model-based belief formation, communication and polarization, and belief persistence. It extends these literature by investigating whether and how non-strategic communication affects both consensus and truth-finding in a common-value narrative environment where participants observe different private data and may interpret those data differently.

First, the paper relates to a recent and fast-growing literature that studies narratives

as mental models or causal structures used to interpret observed facts.[2] A leading contribution is Eliaz and Spiegler (2020), who model narratives as subjective causal structures represented by directed acyclic graphs. In their framework, a narrative selects variables and causal links, and beliefs are formed by applying the Bayesian-network factorization implied by the narrative to objective data. Different narratives can therefore generate different beliefs from the same empirical correlations. Subsequent work extends this causal-narrative approach to political and media environments, studying how narrative competition, strategic narrative provision, and false narratives can sustain disagreement and polarization (Eliaz and Spiegler, 2024; Eliaz et al., 2025). This framework has also motivated a growing experimental literature.[3] In particular, Charles and Kendall (2023) and Ambuehl and Thysen (2025) test behavioral implications of causal-narrative models and document heterogeneity in how individuals evaluate and rely on competing causal explanations. Our paper shares with this literature the view that narratives are interpretive devices. We differ by studying how disagreement over narratives evolves through interpersonal communication, while holding fixed the set of admissible narratives and the underlying true data-generating process.

A complementary approach models narratives as likelihood-based models that organize data and guide prediction. Schwartzstein and Sunderam (2021) develop a model-persuasion framework in which a sender proposes alternative models to help a receiver interpret past data. Receivers find models more compelling when they better fit the data, but better-fitting models leave less room for belief distortion. Aina (2023) builds on this approach and studies the extent to which a biased sender can manipulate posterior beliefs by tailoring the set of models available to the receiver. Barron and Fries (2025) provides experimental evidence that narratives are persuasive and that narrative fit is an important determinant of persuasion. These papers show that communication about models can shape beliefs even when the objective data are held fixed.

More recent work studies environments in which agents exchange narratives or face

[2]This line of work builds on a broader literature on narratives in economics, pioneered by Shiller (2017, 2020), which emphasizes the macroeconomic and social roles of narratives more generally. For a recent survey of the broader narratives literature, see Roos and Reccius (2024).

[3]Within the broader narratives literature, several experimental studies examine the effects of narratives on beliefs and behavior across different domains (Yang and Hobbs, 2020; Harrs et al., 2021; Barron et al., 2023).

competing models. Schwartzstein and Sunderam (2025) examine how agents share models or interpretations to make sense of new data in a community. Montiel Olea et al. (2022) study agents who observe common data but use different predictive models, and show how perceived predictive performance can sustain disagreement. A related experimental contribution is Liu and Zhang (2025), who study whether later counter-narratives can undo the effects of earlier narrative exposure. Our paper differs from these contributions in two ways. First, we abstract from strategic persuasion, as participants have aligned incentives to identify the true narrative. Second, agents observe different private samples. Communication can therefore serve two distinct functions in our setting: exchanging data and exchanging interpretations, both of which may facilitate truth-finding. This dual role is central to the experiment and to many real-world disputes over factual issues.

Second, our paper speaks to an experimental literature on how individuals form mental models from data. Kendall and Oprea (2024) study how subjects form predictive models of simple data-generating processes and show that subjects often struggle to form or explicitly describe the model they use. Fréchette et al. (2024) study how individuals infer statistical relationships from observational data and document substantial heterogeneity in the models subjects extract. Aina and Schneider (2025) provide evidence on how individuals weight competing models, including cases in which they place most weight on the best-fitting model rather than averaging across models in a fully Bayesian manner. Related experimental work documents systematic errors arising from misspecified mental models (Hanna et al., 2014; Enke, 2020; Esponda et al., 2024), showing that individuals may neglect relevant dimensions of the data or rely on incomplete representations of the environment. These studies focus primarily on individual belief formation. We extend this line of work to an interactive environment in which individuals communicate both their observed evidence and their reasoning, and in which the central outcomes are whether communication generates consensus and improves truth-finding.

Third, this paper relates to the empirical literature on communication, social interaction, and polarization. Braghieri et al. (2025) study cross-partisan conversations between Democrats and Republicans and show that such interactions are perceived as uninformative ex-ante, even though they may reduce affective polarization ex post. Santoro and Broockman (2022) provide large-scale experimental evidence that structured cross-partisan conversations can meaningfully reduce affective polarization, while having

more limited effects on substantive belief convergence. Similarly, Fang et al. (2025) study a large-scale initiative in Germany and find that conversations across political divides reduce affective polarization but need not generate convergence in ideological views. Schwardmann et al. (2022) show that being required to argue for a position can lead individuals to persuade themselves that the position is correct, making later belief revision more difficult. These papers study communication in politically or socially charged environments, where identity, ideology, and strategic self-persuasion motives may be central. Our experiment studies a related but distinct problem: how communication affects disagreement over the interpretation of observed facts when incentives are aligned and the true state is commonly valued. This allows us to isolate interpretive frictions from partisan identity and strategic persuasion.

Finally, our findings contribute to the literature on confirmation bias and commitment bias (Rabin and Schrag, 1999; Kahneman et al., 1990; Staw, 1976). While this literature documents many settings in which individuals adhere to existing beliefs or choices by selectively interpreting information, we provide suggestive evidence that encouraging individuals to articulate their reasoning makes them less likely to change their minds even when an objective truth exists. This result is distinctive because our setting features a single underlying truth and non-strategic communication, closely mirroring everyday conversations. Our findings also relate to the concept of "beliefs as assets" (Bénabou and Tirole, 2011), as beliefs in our setting take the form of sense-making narratives that individuals become attached to and reluctant to revise.

3 Experimental design, procedures, and hypotheses

The experiment has two stages. In Stage 1, each participant observes a private sample of 20 observations generated from one of two possible narratives and then chooses which narrative is true. In Stage 2, participants are randomly paired and then communicate in free-format chatboxes before making the same choice again. We vary the communication environment across three between-subject treatments: Communication, Share, and Reasoning. The main outcomes of interest are whether paired participants reach consensus and whether their choices correspond to the true or best-fitting narrative.

3.1 Narrative setting and data-generating process

The experiment studies how people form and communicate narratives when there is an objectively correct data-generating process while observing only finite and noisy samples. We use explanations of the gender wage gap in China as a narrative domain. This setting is useful for three reasons. First, the topic is familiar and salient for our participant pool of Chinese college students, yet there is no clear consensus among them about the underlying explanation. Second, within the experiment, the setting admits a well-defined true narrative. Third, identifying the underlying narrative from observing a finite dataset is neither trivial nor mechanical.

The gender wage gap refers to the average wage difference between male and female workers. Empirical studies commonly decompose this gap into an *explained component* and an *unexplained component*. The explained component captures the part of the gap attributable to gender differences in observable characteristics, such as education, work experience, working hours, or productivity. The unexplained component is the residual gap after accounting for such observable characteristics and is often interpreted as a proxy for unequal treatment or discrimination.[4]

3.1.1 Two candidate narratives

Participants choose between two narratives about the decomposition of the gender wage gap. Both narratives are grounded in empirical studies using recent Chinese data. Because these studies yield markedly different decomposition results, we adopt their respective estimates to construct two distinct data-generating processes.

The first narrative follows the estimate in Zhang et al. (2023), according to which the unexplained component accounts for more than 90% of the gender wage gap.[5]

Narrative 1. *Less than 10% of the gender wage gap is attributed to the explained component, while more than 90% is attributed to the unexplained component.*

The second narrative follows the estimate in Ma (2022), in which the explained and

[4]See Iwasaki and Ma (2020) for a survey of empirical studies on the gender wage gap in China. Survey evidence also documents widespread public discussion and divergent attitudes toward gender inequality, particularly in the workplace (Wang et al., 2024; Ma, 2025).

[5]Zhang et al. (2023) estimate that the unexplained component accounts for 94.72% of the gender wage gap. For ease of comprehension in the experiment, we state this as "more than 90%."

unexplained components each account for about 50% of the gender wage gap.[6]

Narrative 2. *The explained component and the unexplained component each account for approximately 50% of the gender wage gap.*

At the beginning of the experiment, the computer randomly assigns one of these two narratives as the *true narrative*, with equal probability, for all participants within the same matching group (see Section 3.3 for further description). The assigned true narrative determines the data pool from which participants' private graphs are sampled.

3.1.2 Data-generating process and personal data graphs

For each true narrative, we construct a dataset that contains 1,000 observations. Each observation consists of wage, productivity, and gender. The two datasets are calibrated so that their wage–productivity relationships reproduce the empirical patterns implied by Narrative 1 and Narrative 2, respectively. Appendix A describes the data-generating process, and Figure A.1 presents the two datasets.

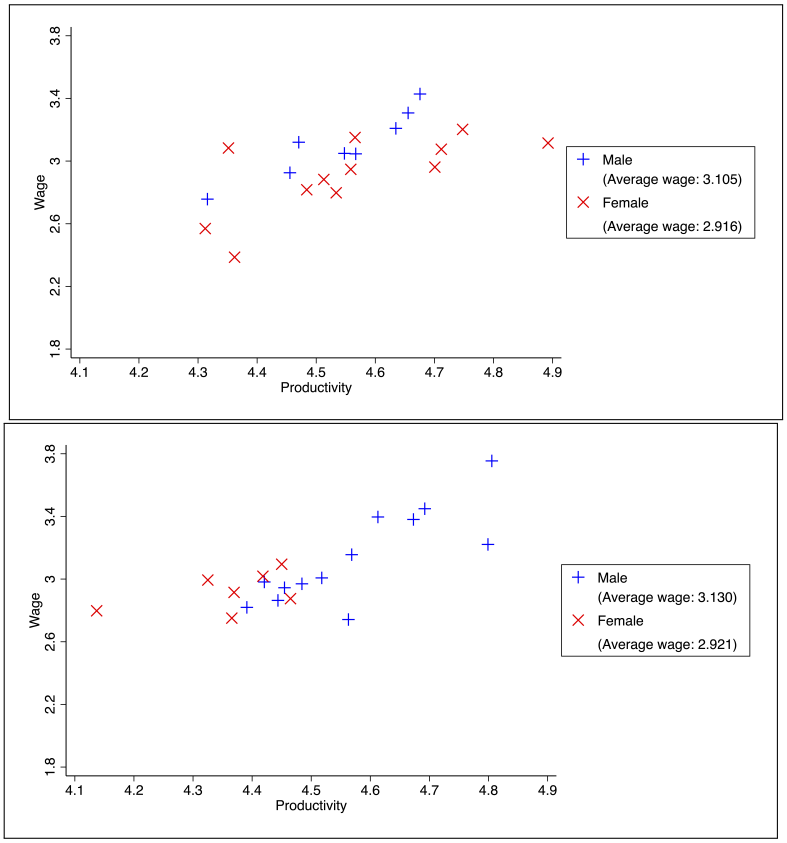
Conditional on the assigned true narrative, the computer randomly and independently draws 20 observations from the corresponding dataset for each participant.[7] The 20 observations are displayed as a data graph. The horizontal axis reports productivity and the vertical axis reports wage; male and female observations are represented by different markers. The graph also reports the average wage of male and female observations in the participant's sample. Figure 1 shows examples of personal data graphs generated under Narrative 1 and Narrative 2. In the experimental instructions, participants are informed that the true narrative is randomly selected with equal probability, and learn about the main features of the data-generating process and the construction of their personal data graphs, but are not informed of all details of the data-generating process described in Appendix A, including the specific distributional assumptions used to generate wages and productivity.[8]

[6] Ma (2022) reports that the unexplained component accounts for 51.2% of the gender wage gap. We round this to "approximately 50%." The two studies are comparable in that both use the same decomposition method and analyze Chinese data spanning multiple sectors, regions, and age groups from 2014–2018. The main difference lies in the data sources used by the two studies.

[7] Because observations are drawn randomly, the number of male and female observations in a given participant's sample need not be equal.

[8] Appendix B provides the experimental instructions and screenshots.

Figure 1: Examples of data graphs when the true narrative is Narrative 1 (top panel) or Narrative 2 (bottom panel).



3.2 Experimental design and treatments

We implement a two-stage experimental design.

3.2.1 Stage 1: Individual baseline

Stage 1 provides the individual baseline. Each participant observes only their own 20-observation data graph and then chooses which of the two narratives they believe to be the true data-generating process. A correct choice pays 40 Chinese yuan (CNY), while an incorrect choice pays zero. This stage is selected for payment with probability 50%.

After making Stage 1 choices, participants within the same matching group are then randomly matched into pairs. Based on their Stage 1 choices, a pair is classified as a *same-narrative pair* if both members initially choose the same narrative, and as a *conflicting-narrative pair* if they initially choose different narratives.

Participants do not receive feedback after their choices in Stage 1. Thus, when they enter Stage 2, they only know their own initial narrative choices and personal data graphs,

but not whether their initial choices are correct.

3.2.2 Stage 2: Paired Communication

In Stage 2, paired participants communicate through an anonymous computer-based chat interface. Communication lasts for up to 15 minutes and may end earlier if both participants agree to stop.[9] After communication, participants are requested to choose between Narrative 1 and Narrative 2 again. A correct choice in Stage 2 also pays 40 CNY. At the end of the experiment, one of the two narrative-choice stages is randomly selected for payment with equal probability.

We implement three between-subject treatments. These treatments introduce variations in whether participants can directly observe their partner's data and whether they are prompted to articulate their reasoning before communication.

The *Communication* treatment. During the chat of this treatment, each participant observes both players' Stage 1 narrative choices and can revisit her own private data graphs. Participants therefore know whether their initial narrative choices agree or not. However, they cannot directly observe their partner's data graph. Any information about the partner's data must then be transmitted through free-form communication.

This treatment captures a common feature of real-world narrative exchange, where individuals often know others' views but do not directly observe the evidence or reasoning that generated those views. Communication may therefore operate through two potentially imperfect channels, describing private data and explaining how those data are interpreted.

The *Share* treatment. This treatment introduces data sharing. In contrast to *Communication*, during the chat each participant can directly observe her partner's data graph, as well as a combined graph showing the 40 observations from both members of the pair. Relative to *Communication*, *Share* removes the need to exchange private data verbally. Thus, it should improve consensus and correctness if disagreement reflects imperfect data transmission, but have limited effects if disagreement reflects differences in data interpretation.

The *Reasoning* treatment. The *Reasoning* treatment tests the effect of prompting

[9]Participants are explicitly informed that the two members of the pair face the same true narrative and that this true narrative remains fixed across the two stages.

participants to articulate their reasoning before communication. It is identical to *Communication* during the chat, but adds an advice-writing stage before communication. After making the Stage 1 choice and before communicating with the matched partner, each participant writes an advice of up to 250 words for a future participant on how to infer the true narrative from the privately observed data graph. Participants are told that the future participant will face the same true narrative but may observe a different personal data graph. While writing the advice, participants can review their own data graph and Stage 1 choice.

Advice is incentivized. In a follow-up evaluation task, each future participant reads three pieces of advice and selects the most helpful one. If a participant's advice is selected, the participant receives 20 CNY; otherwise, she receives zero. After writing advice, participants enter the same communication environment as in *Communication*.

Ex ante, this treatment may have opposing effects. Writing advice may improve communication by helping participants organize and articulate their inference strategy. At the same time, articulating a rationale for one's initial choice may increase attachment to that choice and to the reasoning behind it.[10] This treatment therefore identifies the effect of reasoning before communication.

3.3 Procedures

The main experiment was conducted at the Shanghai University of Finance and Economics in April 2025. Participants were recruited from the Economic Lab and took part in only one treatment.

We conducted 13 sessions, each with three or four matching groups of 6–10 participants. The three between-subject treatments were randomized at the session level, with each session implementing one treatment. At the beginning of each session, the computer randomly assigned a true narrative to each matching group. At the end of Stage 1, participants were randomly paired within their matching group, ensuring that both participants in a pair faced the same true narrative.[11] Table 1 reports the numbers of participants, pairs, conflicting-narrative pairs, and same-narrative pairs by treatment.

[10] We discuss several behavioral mechanisms that could generate this effect in Section 3.4 below.

[11] Thus, the independent observations are individuals for Stage 1 choices and pairs for Stage 2 choices. Matching groups determine the assigned true narrative but do not otherwise affect independence.

In total, 342 participants participated in the main experiment: 112 in *Communication*, 110 in *Share*, and 120 in *Reasoning*.

Table 1: Summary of participants by treatments

	No. of participants	No. of pairs	No. of conflicting-narrative pairs	No. of same-narrative pairs
Communication	112	56	21	35
Share	110	55	26	29
Reasoning	120	60	23	37

The experiment was programmed in z-Tree (Fischbacher, 2007). Upon arrival, participants were randomly seated at computer terminals. At the beginning of each stage, they read the relevant instructions on screen and had to answer all control questions correctly before proceeding. Communication was anonymous and conducted only through the computer interface; participants had no opportunity to identify their partner.

After the two narrative-choice stages, participants completed post-experimental elicitation and a demographic questionnaire. We elicited their theory of mind (ToM) and cognitive abilities. The detailed elicitation methods are presented in Appendix B.6. At the end of the session, participants were informed which narrative-choice stage had been randomly selected for payment. The average payment in the main experiment was 70 CNY, including a 20 CNY show-up fee. Each session lasted approximately 60 minutes.

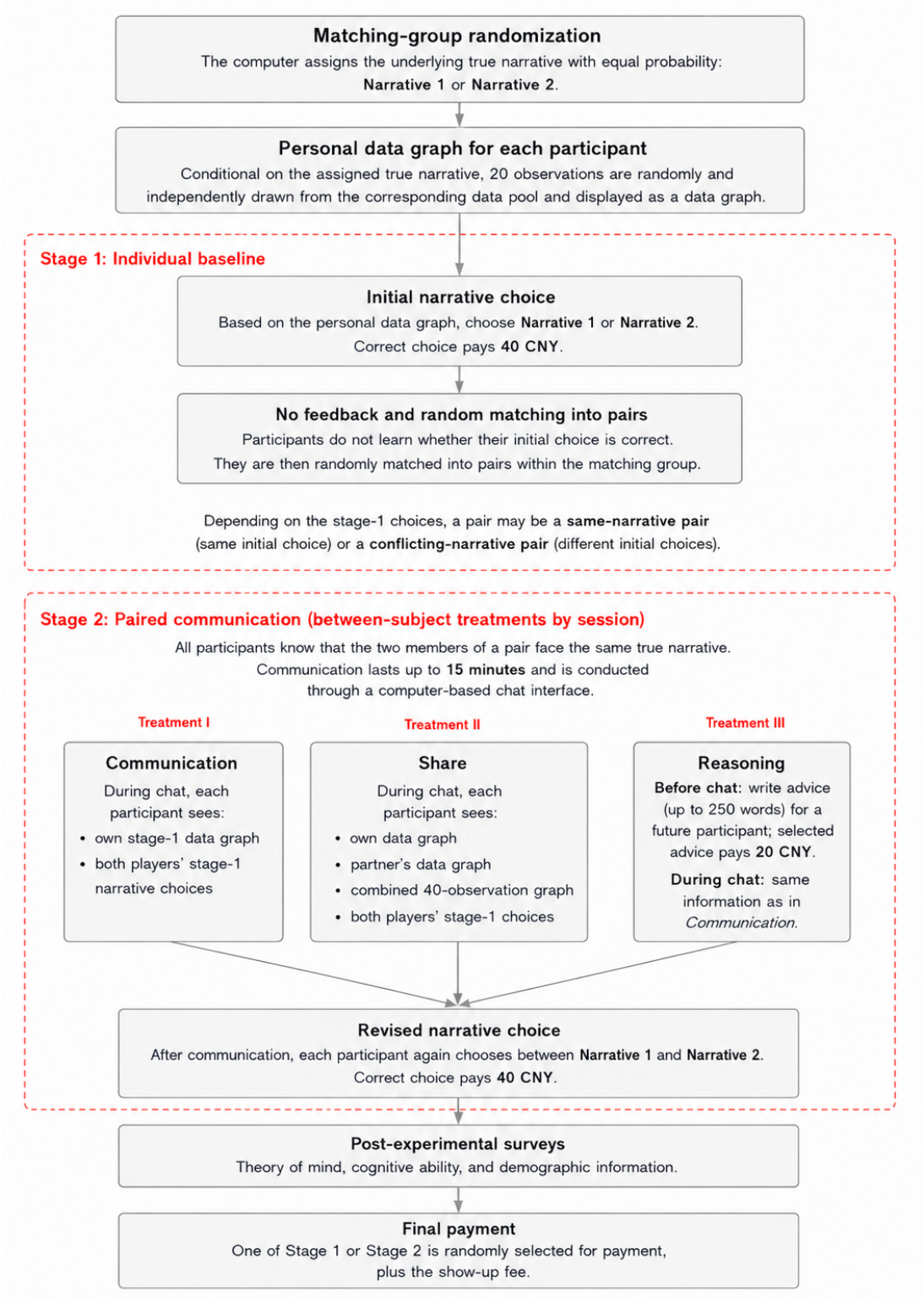
The advice-evaluation follow-up experiment for the *Reasoning* treatment was conducted at another university within two weeks of the main experiment. We recruited 44 participants to evaluate the 120 pieces of advice written by participants in *Reasoning*. The follow-up consisted of two 25-minute sessions, with an average payment of 31 CNY.

Figure 2 illustrates the experimental design and procedure.

3.4 Hypotheses

We focus on two main outcomes: consensus and correctness. Consensus is measured at the pair level and equals one if both members of a pair choose the same narrative in Stage 2. We calculate the consensus rate separately for the same-narrative and conflicting-narrative pairs, with a primary emphasis on the conflicting-narrative pairs because these pairs begin Stage 2 without agreement.

Figure 2: Experimental design and procedure



Correctness is measured at the individual level. The most direct measure is whether the participant's narrative choice matches the true narrative assigned by the computer. Because participants observe finite and noisy samples, we also consider in the analysis a complementary *best-fit* measure: the narrative that best fits the data available to the participant, either her personal data or the pair's combined data. The true-narrative measure captures objective truth-finding, while the best-fit measure captures whether the participant chooses the narrative better supported by observable evidence.

We compare *Communication* with *Share* to identify the role of data sharing, which effectively eliminates information asymmetry in private observations. In a conflicting-narrative pair under *Communication*, consensus requires participants to perform two tasks: exchange and integrate information about different private samples, and agree on how to draw the final conclusion from those samples. *Share* removes the first task by directly displaying both the personal data graphs and the combined graph. This leads to the first hypothesis about consensus.

Hypothesis 1 (Consensus, *Communication* vs. *Share*)**.** *Among conflicting-narrative pairs, the consensus rate is higher in Share than in Communication.*

The same logic suggests that direct data sharing should improve the correctness rate. This is because the combined 40-observation graph contains less sampling noise than either 20-observation personal graph. Holding fixed the mental model used to interpret the graph, participants should therefore be more likely to identify the correct or best-fit narrative in *Share* than in *Communication*.

Hypothesis 2 (Correctness, *Communication* vs. *Share*)**.** *Correctness rates are higher in Share than in Communication.*

We next compare *Communication* with *Reasoning*. A priori, prompting participants to articulate their reasoning process can have an ambiguous effect on consensus formation. On the one hand, if the main barrier to consensus is that participants find it difficult to explain their inference strategy during the chat, then the advice-writing task should improve communication. By making participants formulate their reasoning in advance, the task may help them communicate more clearly and evaluate their partner's arguments more effectively. This mechanism predicts higher consensus in *Reasoning*.

Hypothesis 3a (Consensus, *Communication* vs. *Reasoning*)**.** *Among conflicting-narrative pairs, the consensus rate is higher in Reasoning than in Communication.*

However, the advice-writing task can also reduce the probability of consensus. Once participants explicitly articulate a rationale for their Stage 1 choice, they can become more committed to both the chosen narrative and the mental model that supports it. Several behavioral mechanisms can generate such an effect. First, explicit articulation can strengthen confirmation bias: participants can subsequently interpret their partner's information in a way that favors the narrative they have already justified (Rabin and Schrag, 1999). Second, it may create commitment bias: after giving advice and articulating reasoning to others, participants may become reluctant to abandon their earlier choice even when exposed to conflicting evidence (Staw, 1976). Third, constructing an argument may induce an attachment to one's own rationale: once a rationale is perceived as one's own, participants may attach additional value to it and become less willing to give it up, in the spirit of the endowment effect (Kahneman et al., 1990). Finally, the advice task may induce self-persuasion. As found in Schwardmann et al. (2022), the act of preparing an argument for one's conclusion can lead individuals to become more convinced of that conclusion. In our setting, these forces may make participants' initial narratives more resistant to revision, especially when their partner begins from a conflicting narrative. This leads to the competing consensus hypothesis.

Hypothesis 3b (Consensus, *Communication* vs. *Reasoning*)**.** *Among conflicting-narrative pairs, the consensus rate is lower in Reasoning than in Communication.*

Similarly, the reflection induced by advice writing can potentially improve correctness by facilitating more careful analysis of the data and a more logical mapping from data to narratives. This mechanism predicts that participants in *Reasoning* are more likely to choose the true narrative, or at least the narrative better supported by the available data.

Hypothesis 4 (Correctness, *Communication* vs. *Reasoning*)**.** *Correctness rates are higher in Reasoning than in Communication.*

Finally, because *Share* and *Reasoning* differ along two dimensions, we do not formulate a direct hypothesis comparing them.

4 Experimental Results

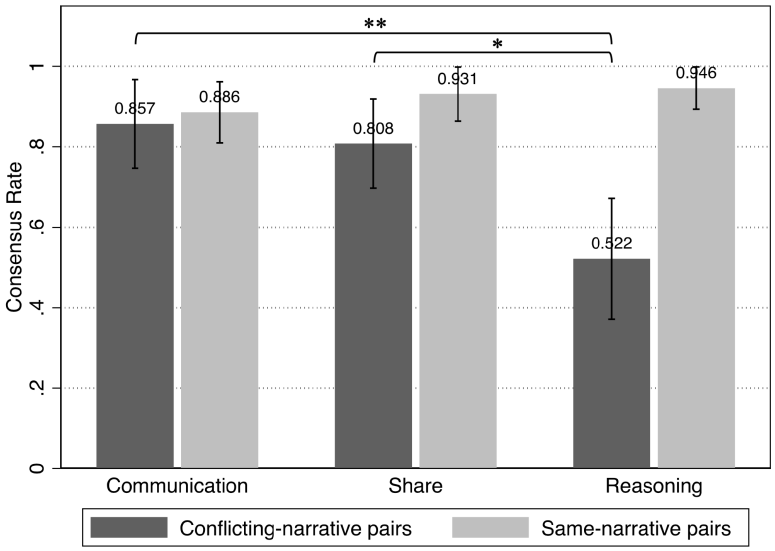
4.1 Treatment effects on aggregate outcomes

This subsection reports the treatment effects on two main aggregate outcomes: consensus and correctness. Consensus captures whether communication leads the two members of a pair to select the same narrative in Stage 2. Correctness captures whether communication improves the quality of narrative choices. We use two notions of correctness. *True-narrative correctness* equals one if the choice matches the true narrative that is used to generate the data graph. *Best-fit correctness* equals one if the choice matches the narrative that best fits the participant's available data. This distinction matters because participants observe finite and noisy samples, so the narrative best supported by realized data may differ from the underlying true data-generating process.

4.1.1 Consensus

We begin with consensus. Table 1 reports the number of same-narrative and conflicting-narrative pairs in each treatment. Figure 3 reports the Stage 2 consensus rate separately for the two types of pairs.

Figure 3: Consensus rate by treatment and pair type



Among same-narrative pairs, consensus remains high in all three treatments. The Stage 2 consensus rate is 89% in *Communication*, 93% in *Share*, and 95% in *Reasoning*. These differences are not statistically significant, as expected, since these pairs already

chose the same narrative in Stage 1. Consensus is nevertheless not mechanical because a small fraction of initially aligned pairs diverge after communication, indicating that communication can still induce belief revision.

The more important comparison concerns conflicting-narrative pairs. In *Communication*, 86% of initially disagreeing pairs reach consensus after communication. Direct data sharing does not increase this rate: the consensus rate is 81% in *Share*, statistically indistinguishable from *Communication* (two-sided Mann–Whitney test, $p = 0.962$). By contrast, the consensus rate falls to 52% in *Reasoning*. This rate is significantly lower than in *Communication* and marginally lower than in *Share* (two-sided Mann–Whitney tests, $p = 0.037$ and $p = 0.067$, respectively). Thus, direct access to the partner's data does not appear to make consensus more likely, whereas prompting participants to articulate their reasoning before communication substantially reduces consensus among initially disagreeing pairs.

These findings do not support Hypothesis 1, which predicts higher consensus in *Share* than in *Communication*, nor Hypothesis 3a, which predicts higher consensus in *Reasoning*. Instead, the evidence is consistent with Hypothesis 3b: advice writing appears to make participants less willing to abandon their initial narrative when facing disagreement. The aggregate consensus result therefore suggests that, instead of facilitating consensus, articulating reasoning explicitly makes belief revision harder, possibly because it strengthens attachment to one's own interpretation.

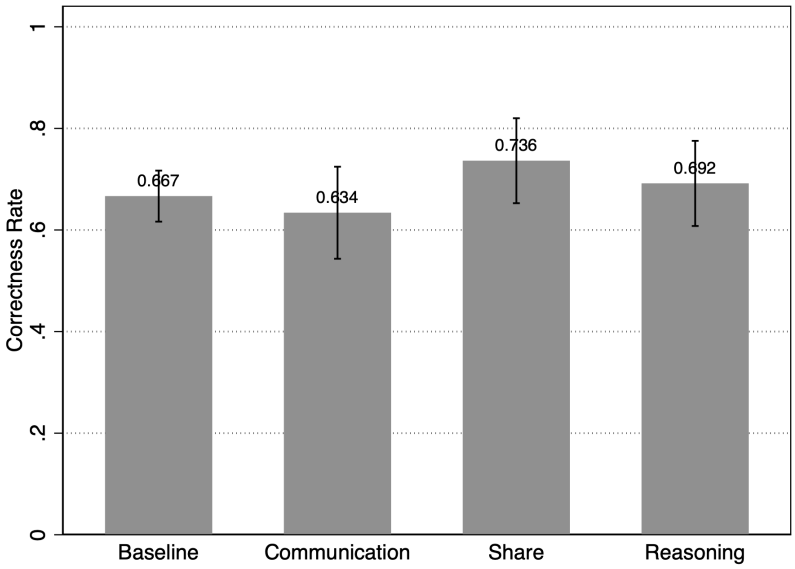
Result 1 (Consensus)**.** *Among conflicting-narrative pairs, the consensus rate is high in both Communication and Share, with no significant difference between the two treatments. The consensus rate is substantially lower in Reasoning, consistent with Hypothesis 3b.*

4.1.2 Correctness relative to the true narrative

We next examine whether communication improves true-narrative correctness. Figure 4 reports the fraction of participants choosing the true narrative in the individual baseline and after communication in each treatment.

The baseline true-narrative correctness rate is 67%. After communication, it is 63% in *Communication*, 74% in *Share*, and 69% in *Reasoning*. None of the pairwise treatment differences is statistically significant, nor does the correctness rate change significantly

Figure 4: True-narrative correctness rate by treatment



from Stage 1 to Stage 2 within any treatment.[12] Thus, although communication often generates consensus, it does not significantly increase the probability of choosing the objectively true narrative. Consensus and truth-finding are therefore empirically distinct, as pairs may converge on a common narrative without systematically moving toward the true data-generating process.

The absence of significant treatment effects on true-narrative correctness indicates that neither direct data sharing nor pre-communication reasoning clearly improves objective truth-finding in this finite-sample environment. In *Share*, the point estimate is higher than in *Communication*, but not significantly so. In *Reasoning*, which most strongly affects consensus, true-narrative correctness rate remains close to the baseline. Hence, Hypotheses 2 and 4 are not supported when correctness is defined relative to the underlying true narrative.

4.1.3 Correctness relative to the best-fit narrative

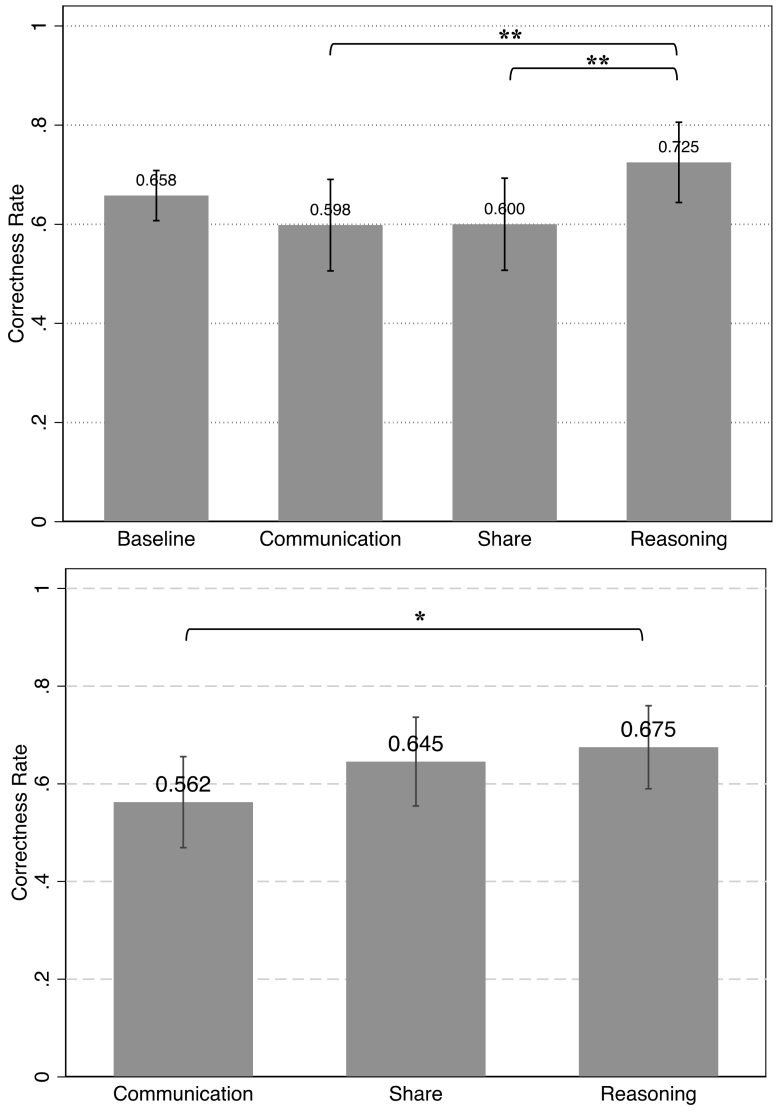
True-narrative correctness is a demanding benchmark because participants observe only finite random samples. A participant may therefore choose the narrative best supported

[12]Two-sided Mann–Whitney tests for pairwise comparisons yield *Communication* vs. *Share*, $p = 0.455$; *Communication* vs. *Reasoning*, $p = 0.353$; *Reasoning* vs. *Share*, $p = 0.101$. The corresponding tests comparing Stage 1 and Stage 2 within each treatment yield $p = 0.172$, $p = 0.526$, and $p = 0.616$, respectively.

by her realized data but still be incorrect relative to the true data-generating process. In light of this possibility, we also evaluate choices using best-fit correctness. For each personal data graph and each pair's combined data graph, we identify the narrative that better fits the observed data. Specifically, we estimate two linear regressions of wage on productivity and a gender dummy, fixing the gender coefficient at the value implied by Narrative 1 or Narrative 2. The narrative associated with the smaller sum of squared errors is classified as the best-fit narrative. Appendix D.1 provides the details.

Figure 5 reports best-fit correctness rates. The top panel uses each participant's personal data graph to construct the best-fit narrative. The bottom panel uses the pair's combined data graph; this measure is considered only for stage 2.

Figure 5: Best-fit correctness rate based on personal data (top panel) and group data (bottom panel)



The results differ sharply from those based on the true narrative. When best-fit correctness is measured using personal data, the rate is 60% in both *Communication* and *Share*, but rises to 73% in *Reasoning*. The difference between *Reasoning* and each of the other two treatments is statistically significant (two-sided Mann–Whitney tests, $p = 0.042$ and $p = 0.045$, respectively). The baseline rate is 66%, which does not differ significantly from any communication treatment (two-sided Mann–Whitney tests, $p = 0.785$, $p = 0.488$, and $p = 0.775$, respectively). Hence, *Reasoning* improves participants' ability to choose the narrative that best fits their own observed data.

The group-data benchmark shows a similar but weaker pattern. Best-fit correctness based on the combined data graph is 56% in *Communication*, 65% in *Share*, and 68% in *Reasoning*. The difference between *Reasoning* and *Communication* is marginally significant (two-sided Mann–Whitney test, $p = 0.078$), while the difference between *Reasoning* and *Share* is not ($p = 0.637$). This suggests that the accuracy gain in *Reasoning* reflects more careful interpretation of the available evidence induced by advice giving.

Together, the aggregate results reveal a trade-off. *Reasoning* makes initially disagreeing participants less likely to reach consensus but more likely to choose the narrative best supported by their data. In contrast, *Share* removes data-transmission frictions but does not improve consensus or best-fit correctness relative to *Communication*. Thus, communication can generate consensus without improving truth-finding, while explicit reasoning can improve evidence-based choices while making narratives harder to revise.

Result 2 (Correctness)**.** *Correctness relative to the true narrative does not differ significantly across treatments. In contrast, best-fit correctness is significantly higher in Reasoning than in both Communication and Share when evaluated using personal data, and remains higher than Communication when evaluated using group data. These findings support Hypothesis 4 for best-fit correctness, but do not support Hypothesis 2.*

4.2 Mechanisms behind the aggregate treatment effects

The aggregate results in Section 4.1 reveal a trade-off: *Reasoning* substantially lowers consensus among conflicting-narrative pairs, but increases the likelihood that participants choose the narrative best supported by their available data. This subsection investigates the individual-level behavior underlying this pattern.

We organize the analysis around three mechanisms. First, we examine whether

participants exhibit a bias toward their own private data when making narrative choices. Second, among conflicting-narrative pairs that reach consensus, we investigate whether stronger evidence has a greater influence on the agreed-upon narrative. Third, we examine whether initial disagreement induces participants to revise toward the true narrative relative to initially aligned pairs.

4.2.1 Measuring best-fit narratives

Our analysis relies on an empirical measure of which narrative is better supported by a given data graph. We use the same best-fit measure introduced in Section 4.1 and detailed in Appendix D.1. For any dataset D — either a participant's private 20-observation dataset or a pair's combined 40-observation dataset — we evaluate the fitness of Narrative 1 and Narrative 2 using restricted linear regressions. Specifically, for each narrative $k \in \{1, 2\}$, we estimate

$$Y_\ell = \alpha_k + \beta_k X_\ell + \gamma_k G_\ell + \varepsilon_\ell, \qquad \ell \in D,$$

where Y_ℓ is worker ℓ's wage, X_ℓ is worker ℓ's productivity, G_ℓ indicates whether worker ℓ is male. The gender dummy coefficient γ_k is fixed at the value implied by the relative importance of the unexplained component implied by Narrative k, while α_k and β_k are chosen to minimize the sum of squared errors. Let $SSE_k(D)$ denote the resulting residual sum of squares. Narrative 1 is the best-fit narrative for D if $SSE_1(D) < SSE_2(D)$; otherwise, Narrative 2 is the best-fit narrative. Intuitively, the best-fit narrative is the one whose implied decomposition better explains the observed wage-productivity pattern.

This construction gives three objects used below. First, each participant has an *own best-fit narrative*, based on her own data graph. Second, in *Share*, the *group best-fit narrative*, based on the combined 40-observation data graph, also becomes relevant. Third, when a conflicting-narrative pair reaches consensus, we ask whether the pair converges to the narrative supported by the stronger individual evidence. Let $SSE_i^* = \min\{SSE_1(D_i), SSE_2(D_i)\}$ denote the best attainable SSE given participant i's own private dataset. Within a pair, the participant with the lower SSE_i^* is classified as having stronger-fitting personal evidence, and her own best-fit narrative is defined as the pair's *stronger-fit narrative*.

Table 2: Reliance on own best-fit narrative in Stage 2

	Treatment		
	Communication	Share	Reasoning
Panel A. For all participants whose own best-fit narrative differs from partner's:			
Proportion	0.500 (0.506)	0.500 (0.506)	0.667 (0.478)
T-test p-value (H_0: Proportion = 0.5)	1.000	1.000	0.044**
No. of observations	44	42	36
Panel B. For participants in conflicting-narrative pairs whose own best-fit narrative differs from partner's:			
Proportion	0.563 (0.512)	0.571 (0.514)	0.813 (0.403)
T-test p-value (H_0: Proportion = 0.5)	0.633	0.612	0.007***
No. of observations	18	14	16
Panel C. For all participants whose own best-fit narrative differs from pair's:			
Proportion	0.591 (0.503)	0.381 (0.498)	0.667 (0.485)
T-test p-value (H_0: Proportion = 0.5)	0.406	0.286	0.163
No. of observations	22	21	18
Panel D. For participants in conflicting-narrative pairs whose own best-fit narrative differs from pair's:			
Proportion	0.750 (0.463)	0.429 (0.535)	0.875 (0.354)
T-test p-value (H_0: Proportion = 0.5)	0.171	0.736	0.020**
No. of observations	8	7	8

Notes: Standard deviations are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.2.2 Bias toward private evidence

We first explore whether participants exhibit a bias toward their personal data; that is, whether they rely more on their private data than on their partner's or the group's data. A direct way to assess this is to examine choices when a participant's own best-fit narrative conflicts with her partner's. Table 2 reports the share of participants choosing their own best-fit narrative in Stage 2, restricting attention to cases where it differs from either the partner's own best-fit narrative or the pair's group best-fit narrative.

The evidence is strongest for *Reasoning*. When participants' own best-fit narrative differs from their partners', those in *Reasoning* choose their own best-fit narratives 66.7% of the time in the full sample and 81.3% of the time among conflicting-narrative pairs; both rates are significantly above 50%. The pattern is even stronger when their own best-fit narrative differs from the group best-fit narrative. Specifically, among conflicting-narrative pairs, 87.5% choose the narrative best supported by their own data. This

suggests that articulating one's reasoning increases reliance on her private evidence.

By contrast, participants in *Communication* do not significantly favor their own best-fit narrative when it differs from their partner's. They choose their own best-fit narrative relatively often when it differs from the group best-fit narrative, but the group data are not directly observed in this treatment. This pattern may therefore reflect participants' difficulty in reconstructing group-level evidence through communication alone.

The pattern in *Share* is different. When participants' own best-fit narrative differs from their partner's, they are about equally likely to choose either narrative. When it differs from the group best-fit narrative, the share choosing the own best-fit narrative falls from 50% to 38.1% in the full sample and to 42.9% among conflicting-narrative pairs. Although these estimates are not statistically significant, they suggest that direct access to the combined data graph reduces the tendency to overweight personal evidence; if anything, participants place relatively more weight on group evidence.

Next, we ask whether conflict in private evidence makes consensus less likely, especially in *Reasoning*. Even among conflicting-narrative pairs, participants' personal data may support the same best-fit narrative. If *Reasoning* increases reliance on private evidence, consensus should be especially difficult when participants' own best-fit narratives point to different directions.

Table 3 provides suggestive evidence for this. In *Reasoning*, conflicting-narrative pairs reach consensus in only 37.5% of the time when their own best-fit narratives differ, compared with 60.0% of the time when they align. This difference is not statistically significant, likely due to the small number of conflicting-narrative pairs, but its direction is consistent with stronger reliance on private evidence. By contrast, the same comparison is essentially flat in *Communication* and *Share*. Thus, while personal best-fit disagreement does not statistically explain consensus formation, the pattern suggests that private evidence plays a larger role in *Reasoning*.

Result 3 (Bias toward personal or group evidence)**.** *Participants in Reasoning are more likely to choose their own best-fit narrative when it conflicts with their partner's or the group's. In Share, choices are instead better predicted by the group best-fit narrative.*

These findings help rationalize the aggregate treatment effects. In *Reasoning*, participants deliberate more carefully but become more anchored to the narrative supported by their own evidence. This improves individual best-fit correctness but reduces

Table 3: Consensus rates by alignment of own best-fit narratives

	Treatment			
	All	Communication	Share	Reasoning
Pairs whose own best-fit narratives differ	0.696	0.875	0.857	0.375
	(0.465)	(0.342)	(0.363)	(0.500)
Pairs whose own best-fit narratives align	0.745	0.846	0.789	0.600
	(0.438)	(0.368)	(0.413)	(0.498)
Mann–Whitney test (p-value)	0.872	1.000	1.000	0.556

Notes: Standard deviations are shown in parentheses. Two-sided Mann–Whitney tests are performed at the pair level ($n = 70$ for All, $n = 42$ for *Communication*, $n = 26$ for *Share*, and $n = 23$ for *Reasoning*). The sample is restricted to conflicting-narrative pairs.

consensus when the two participants' private evidence points to different directions. In *Share*, participants rely more on the combined data graph, which reduces sampling error, but this does not translate into a statistically significant increase in true-narrative correctness.

4.2.3 The role of stronger evidence on consensus

We now explore whether participants with stronger personal evidence — that is, those whose data imply the stronger-fit narrative defined in Section 4.2.1 — are more likely to persuade their partners to revise their narrative choices. Table 4 provides supporting evidence for this. Among conflicting-narrative pairs that reach consensus, 75% of pairs in *Reasoning* converge to the stronger-fit narrative, significantly above 50% at the ten-percent level. By contrast, the corresponding shares are close to random in *Communication* and *Share*, at 55.6% and 52.4%, respectively. Thus, the lower consensus rate in *Reasoning* does not simply reflect unproductive disagreement; rather, consensus appears to be more disciplined by the relative strength of participants' personal evidence.

Result 4 (Effects of stronger evidence)**.** *Conditional on reaching consensus, conflicting-narrative pairs in Reasoning are more likely to converge to the stronger-fit narrative. No comparable pattern appears in Communication or Share.*

This result refines the interpretation of the low consensus rate in *Reasoning*. Prompting participants to articulate their reasoning does not simply make them more stubborn. Instead, it appears to make them more selective: they are less likely to revise

Table 4: Proportion of pairs reaching consensus on the stronger-fit narrative

	Treatment			
	All	Communication	Share	Reasoning
Proportion	0.588	0.556	0.524	0.750
	(0.497)	(0.511)	(0.512)	(0.452)
T-test p-value (H_0: proportion = 0.5)	0.211	0.651	0.833	0.082*
No. of observations	51	18	21	12

Notes: Standard deviations are shown in parentheses. The sample is restricted to conflicting-narrative pairs that reach consensus in Stage 2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

their own best-fit narrative, but when they do reach an agreement, consensus is more likely to favor the narrative with stronger empirical fit.

4.2.4 The role of initial disagreement on truth-finding

Finally, we examine whether initial disagreement facilitates truth-finding. Table 5 reports the change in the share of participants choosing the true narrative from Stage 1 to Stage 2, separately for conflicting- and same-narrative pairs.

Across all treatments, true-narrative correctness increases by 9.3 percentage points among participants in conflicting-narrative pairs, but decreases by 3.0 percentage points among participants in same-narrative pairs. The difference is statistically significant. The same directional pattern appears in all three treatments, with a significant effect in *Share* and a marginally significant effect in *Reasoning*. This suggests that disagreement can induce additional caution and deliberation, whereas initial agreement may reduce incentives to reconsider one's narrative, even when it is incorrect.

Result 5 (Disagreement and revision toward the truth)**.** *Participants in conflicting-narrative pairs are more likely to revise toward the true narrative from Stage 1 to Stage 2 than those in same-narrative pairs. This suggests that disagreement can trigger additional deliberation.*

Overall, the individual-level evidence clarifies the aggregate treatment effects in Section 4.1. *Reasoning* lowers consensus because participants are more likely to stand by the narrative supported by their private evidence; yet, it also makes agreement more selective, such that when consensus occurs, it is more likely to favor the stronger-fit narrative. By contrast, *Share* shifts attention away from personal evidence toward

Table 5: Change in true-narrative correctness between stages

	Treatment			
	All	Communication	Share	Reasoning
Conflicting-narrative pairs	0.093	0.048	0.135	0.087
	(0.645)	(0.661)	(0.687)	(0.590)
Same-narrative pairs	-0.030	-0.029	-0.034	-0.027
	(0.221)	(0.239)	(0.184)	(0.232)
Mann–Whitney test (p-value)	0.010***	0.511	0.046**	0.085*

Notes: Standard deviations are shown in parentheses. Two-sided Mann–Whitney tests are performed at the individual level ($n = 342$ for All, $n = 112$ for *Communication*, $n = 110$ for *Share*, and $n = 120$ for *Reasoning*). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the combined dataset. Finally, initial disagreement has a positive effect, as it induces additional deliberation and improves truth-finding.

4.3 Communication content and treatment effects

The preceding subsection examined how participants revise their narrative choices given the evidence available to them. We now turn to the content of communication. This analysis has two goals: to document how communication differs across treatments, and to examine whether these differences are associated with treatment effects on consensus, narrative revision, and best-fit correctness.

The analysis is exploratory. Chat content is not randomly assigned; it is itself an outcome of the treatment and of participants' endogenous communication choices. The results below should therefore not be interpreted as causal effects of specific conversational behaviors. Instead, they identify communication features that are shifted by the treatment and correlated with behavioral outcomes. We refer to these variables as candidate moderators of treatment effects rather than causal mechanisms.

4.3.1 Coding communication content

Three research assistants independently coded the chat records along multiple dimensions, as described in Appendix C.1. Table C.2 therein defines the coding variables. All variables are binary and recorded at the participant level. When coders disagreed, the final value was determined by majority rule.

Table 6 reports the mean of each variable in the full sample and by treatment; the

Table 6: Summary of coding variables

Variables	Mean				Kruskal-Wallis tests (p-value)
	All	*Communication*	*Share*	*Reasoning*	
same_goal	0.395	0.366	0.327	0.483	0.041**
	(0.490)	(0.484)	(0.471)	(0.502)	
uncertain	0.108	0.045	0.082	0.192	0.001***
	(0.311)	(0.207)	(0.275)	(0.395)	
agree_owndata	0.012	0	0.018	0.017	0.372
	(0.108)	(.)	(0.134)	(0.129)	
diff_data	0.228	0.170	0.227	0.283	0.120
	(0.420)	(0.377)	(0.421)	(0.453)	
describe_data	0.491	0.589	0.209	0.658	<0.001***
	(0.501)	(0.494)	(0.409)	(0.476)	
describe_method	0.716	0.750	0.591	0.800	0.001***
	(0.451)	(0.435)	(0.494)	(0.402)	
describe_method1	0.336	0.357	0.318	0.333	0.826
	(0.473)	(0.481)	(0.468)	(0.473)	
describe_method2	0.117	0.098	0.082	0.167	0.102
	(0.322)	(0.299)	(0.275)	(0.374)	
describe_method3	0.365	0.339	0.291	0.458	0.025**
	(0.482)	(0.476)	(0.456)	(0.500)	
describe_method4	0.094	0.089	0.064	0.125	0.276
	(0.292)	(0.286)	(0.245)	(0.332)	
ask_data	0.132	0.143	0.027	0.217	<0.001***
	(0.339)	(0.351)	(0.164)	(0.414)	
ask_method	0.018	0.018	0	0.033	0.158
	(0.131)	(0.133)	(.)	(0.180)	
real_unexplained	0.012	0.018	0	0.017	0.383
	(0.108)	(0.133)	(.)	(0.129)	
real_explained	0.006	0.009	0.009	0	0.582
	(0.076)	(0.094)	(0.095)	(.)	
persuading	0.015	0.018	0	0.025	0.272
	(0.120)	(0.133)	(.)	(0.157)	
being_persuaded	0.029	0.054	0.009	0.025	0.137
	(0.169)	(0.226)	(0.095)	(0.157)	
state_consensus	0.708	0.732	0.745	0.650	0.223
	(0.456)	(0.445)	(0.438)	(0.479)	

Notes: Standard deviations are shown in parentheses. For Kruskal-Wallis tests, $n = 342$, $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

last column reports Kruskal–Wallis tests of equality across treatments. Six variables differ significantly across treatments. Participants in *Reasoning* are more likely to mention the goal of choosing the same narrative as their partner (*same_goal*), express uncertainty about their initial choice (*uncertain*), describe their data (*describe_data*), describe their reasoning or method (*describe_method* and *describe_method3*), and ask about their partner's data (*ask_data*). The low rates of data description and data requests in *Share* are expected, since both participants' data graphs and the combined graph are directly displayed. By contrast, participants in *Communication* and *Reasoning* must use chat to transmit information about private observations.

Two patterns are especially relevant. First, *Reasoning* generates more explicit discussion of both evidence and reasoning, consistent with the advice task prompting participants to articulate how they infer a narrative from the data. Second, *Share* reduces the need to describe or request data, since the relevant information is already visible to both partners. These align with the evidence in Section 4.2: *Reasoning* increases reliance on articulated private evidence, whereas *Share* shifts attention toward the directly displayed group data.

A caveat is warranted: several variables are rare. In particular, *ask_method*, which captures questions about the partner's mental model, appears in only 1.8% of observations. The two real-world-reference variables, *real_unexplained* and *real_explained*, appear in 1.2% and 0.6% of observations, respectively, and *persuading* occurs in 1.5%. Estimates involving these variables may therefore be driven by a small number of conversations. We report them for completeness, but interpret them as suggestive.

4.3.2 Identifying candidate moderators of treatment effects

To examine whether communication content is related to the treatment effects, we use a two-step diagnostic procedure in the spirit of Braghieri et al. (2025). A communication feature can help explain a treatment difference only if two conditions hold. The treatment changes the frequency of that feature, and conditional on the relevant pre-communication state, the feature predicts the behavioral outcome in the direction needed to account for the treatment effect.

Concretely, for each pairwise treatment comparison and coding variable C_i, the first step tests whether C_i differs across treatments. The second step asks whether

C_i moderates the relevant behavioral transition. For consensus and narrative change, the pre-communication state is membership in a conflicting-narrative pair. For best-fit correctness, it is whether the participant initially fails to choose her own best-fit narrative. The second step therefore estimates whether the association between the pre-communication state and the outcome varies with C_i, using a specification analogous to difference-in-differences. Appendix D.2 reports the corresponding estimates.

We classify a coding variable as a candidate moderator only if it is statistically significant in both steps, using the five-percent level as the benchmark. This deliberately conservative criterion excludes variables that differ across treatments but are unrelated to behavior, as well as variables that predict behavior but are not differentially induced by treatment. This exercise remains correlational. Even when a variable passes both steps, the estimate may reflect omitted conversational features or unobserved participant characteristics rather than the causal effect of that specific content.

Consensus formation. We first apply the procedure to Stage 2 consensus. The aggregate result to be explained is that consensus among conflicting-narrative pairs is substantially lower in *Reasoning* than in the other two treatments. The two-step analysis yields a clear but limited pattern. In the comparison between *Reasoning* and *Communication*, no coding variable passes both steps. Thus, although *Reasoning* lowers consensus relative to *Communication*, the measured content variables do not identify a statistically robust moderator of this difference.

In the comparison between *Reasoning* and *Share*, only two variables pass the two-step criterion: *describe_data* and *describe_method*. Participants in *Reasoning* are more likely to describe both their data and the reasoning behind their choices. These variables are also associated with a lower probability of consensus among initially conflicting pairs. This pattern is consistent with the view that *Reasoning* makes private evidence and reasoning more explicit, but also makes narratives harder to revise when participants begin from different conclusions. In the comparison between *Communication* and *Share*, where aggregate consensus rates are similar, no coding variable emerges as a robust candidate moderator.

Narrative change. The second outcome is whether participants in conflicting-narrative pairs change their narrative choice from Stage 1 to Stage 2. This outcome is closely

related to consensus, since most participants who change narratives in such pairs end up agreeing with their partner. Narrative change is highest in *Share* (48.1%), slightly lower in *Communication* (42.9%), and lowest in *Reasoning* (34.8%), although these aggregate differences are not statistically significant.[13]

The two-step analysis again identifies a candidate moderator only in the *Reasoning–Share* comparison. The variable *describe_method* is more common in *Reasoning* and is negatively associated with narrative change. This reinforces the consensus results, as describing one's method may make the participant's own narrative more coherent and defensible, but it is also associated with a lower willingness to abandon that narrative. No variable passes the criterion in other pairwise comparisons.

Choosing the best-fit narrative. The third outcome is whether participants who initially fail to choose their own best-fit narrative switch to it in Stage 2. This outcome relates to the aggregate finding that *Reasoning* improves best-fit correctness relative to both *Communication* and *Share*.

The clearest candidate moderator that appears in the *Reasoning–Share* comparison is *ask_method*. Participants in *Reasoning* are more likely to ask about the reasoning or method behind their partner's choice, and this variable is positively associated with switching to one's own best-fit narrative. This is consistent with the idea that the advice task not only makes participants articulate their own reasoning, but also makes them more attentive to their partner's reasoning process. Asking about methods may thus help participants update not only their conclusion, but also how they evaluate their data.

This result should be interpreted cautiously. The base rate of *ask_method* is very low, and the variable is absent in *Share*. The estimate may therefore be driven by a small number of conversations in which participants explicitly ask about reasoning. We view this result as suggestive evidence that asking about mental models is associated with better best-fit choices. No robust candidate moderator is identified in other treatment comparisons.

Result 6 (Communication-content moderators)**.** *The chat-content analysis provides suggestive evidence that treatment effects are related to how participants discuss data and*

[13]Two-sided Mann–Whitney tests: *Communication* vs. *Share*, $p = 0.767$, *Communication* vs. *Reasoning*, $p = 0.577$, and *Reasoning* vs. *Share*, $p = 0.260$.

reasoning. Relative to Share, participants in Reasoning more often describe their own data and reasoning; these variables are associated with lower consensus and lower narrative change. Participants in Reasoning are also more likely to ask about their partner's method, which is associated with switching to the best-fit narrative.

Overall, the evidence on the content of the communication complements the individual-level analysis in Section 4.2. *Reasoning* changes the nature of communication by making participants more likely to articulate their data and reasoning. This helps explain why the treatment improves best-fit correctness but reduces consensus among initially conflicting pairs. Making reasoning explicit can improve evidence evaluation, but it can also make initial narratives more resistant to revision. By contrast, *Share* reduces the need to exchange data verbally, and the content analysis provides little evidence that verbal communication in this treatment plays an independent role in improving consensus or correctness.

4.4 Individual traits, data primitives, and narrative choices

The previous subsections examined treatment effects and the behavioral mechanisms that appear to generate them. We now ask a complementary question: which individual characteristics and data-environment primitives predict narrative choices and consensus formation? This exercise serves two purposes. First, it clarifies whether the main treatment effects are partially driven by particular demographic or cognitive groups. Second, it separates the role of communication from the role of the data environment itself, as some participants receive data graphs that make one narrative easier to identify, while others face more ambiguous evidence.

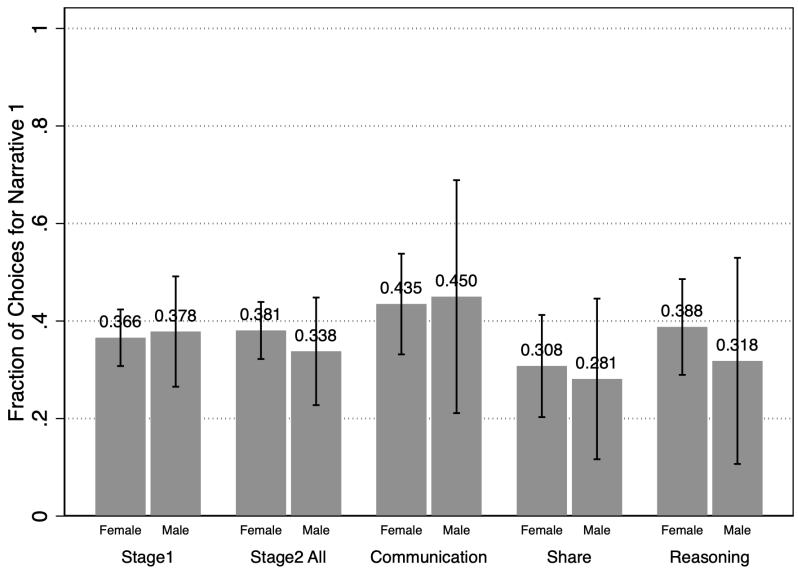
We focus on variables determined before Stage 2 communication. Table C.3 in Appendix C.2 presents their definitions and summary statistics. Individual characteristics include gender, age, theory-of-mind ability, and cognitive ability. Data-environment primitives include the true data-generating process and measures of how clearly the observed data distinguish the two narratives. In particular, *data_own_SSEgap* is the absolute difference between the SSE values of the two narratives evaluated on a participant's own data; larger values indicate that the participant's data more clearly favor one narrative. Similarly, *data_pair_SSEgap* measures the corresponding SSE difference using pair-level data. To capture pair-level disagreement in the data environment, we use

data_ diff_ bestfit, a binary indicator that equals one if the two matched participants' data imply different best-fit narratives and equals zero otherwise.

4.4.1 Gender and narrative selection

We begin with gender effects. Figure 6 reports the proportion of participants choosing Narrative 1 by gender and by stage. In Stage 1, 37% of female participants and 38% of male participants choose Narrative 1. In Stage 2, the corresponding rates are 38% and 34%, respectively. Neither difference is statistically significant (two-sided Mann–Whitney tests, $p = 0.841$ in Stage 1 and $p = 0.501$ in Stage 2). Thus, gender does not predict narrative choice.

Figure 6: Proportions of participants selecting Narrative 1 by gender and by stage



This null result is important because the experiment concerns the gender wage gap. We find no evidence that male and female participants systematically favor different narratives. Both groups choose Narrative 1 less than half of the time in both stages. This pattern may reflect either greater prior plausibility of Narrative 2 or that Narrative 2 is easier to infer from the generated data; our design cannot distinguish between these explanations. The share choosing Narrative 1 also does not differ significantly by gender within any treatment (two-sided Mann–Whitney tests, $p = 1.000$ in *Communication*, $p = 0.974$ in *Share*, and $p = 0.724$ in *Reasoning*). Hence, gender differences do not explain the observed treatment effects.

Table 7: Determinants of choosing the true or best-fit narrative in Stage 1

	(1) True narrative	(2) Own best-fit narrative
Theory-of-Mind Ability	0.007	0.010*
	(0.005)	(0.005)
Cognitive Ability	-0.022	0.015
	(0.021)	(0.022)
Male	0.037	0.080
	(0.059)	(0.061)
Age	0.014	0.042***
	(0.013)	(0.013)
Narrative 2 is true	0.307***	0.124**
	(0.050)	(0.052)
data_own_SSEgap	1.544***	1.235**
	(0.558)	(0.578)
Constant	0.168	-0.499*
	(0.285)	(0.295)
No. of observations	342	342

Notes: Standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.4.2 Stage 1 correctness: individual traits versus data clarity

Table 7 reports linear probability model regressions on Stage 1 choices, before participants communicate with their partners. Column (1) uses true-narrative correctness as the dependent variable, equal to one if the participant chooses the underlying true narrative. Column (2) uses best-fit correctness, equal to one if the participant chooses the narrative that best fits her own personal data. The specification includes individual traits, an indicator for Narrative 2 being true, and the clarity of personal data.

The strongest predictors are data-environment variables. A larger *data_own_SSEgap* significantly increases both true-narrative and own-best-fit correctness. When a participant's data more clearly favor one narrative, she is more likely to choose the correct or best-fitting narrative. The indicator for Narrative 2 being true is also positive and significant in both columns. This mirrors the aggregate pattern in Figure 6, where participants choose Narrative 1 less often overall and are therefore more likely to be correct when Narrative 2 is true. This coefficient should not be interpreted as pure preference for Narrative 2, since it may also reflect differences in the statistical identifiability of the two narratives.

By contrast, individual traits have limited explanatory power. Gender and cognitive

Table 8: Determinants of consensus formation for conflicting-narrative pairs

	(1) All	(2) Communication	(3) Share	(4) Reasoning	(5) All	(6) Communication	(7) Share	(8) Reasoning
Theory-of-Mind Ability	0.001	0.019*	-0.008	0.000	0.002	0.024**	-0.006	-0.002
	(0.007)	(0.011)	(0.010)	(0.019)	(0.007)	(0.010)	(0.011)	(0.019)
Cognitive Ability	-0.017	0.018	-0.012	-0.028	-0.016	0.029	0.003	-0.018
	(0.034)	(0.042)	(0.040)	(0.062)	(0.035)	(0.042)	(0.040)	(0.057)
Male	0.004	-0.088	-0.016	0.041	0.023	0.076	0.006	-0.035
	(0.082)	(0.120)	(0.103)	(0.241)	(0.084)	(0.145)	(0.101)	(0.260)
Age	0.017	0.001	-0.010	0.021	0.019	0.012	-0.003	0.021
	(0.021)	(0.015)	(0.029)	(0.047)	(0.021)	(0.018)	(0.029)	(0.041)
data_ diff_ bestfit	-0.067	0.026	0.076	-0.226	-0.067	-0.053	0.105	-0.239
	(0.123)	(0.162)	(0.170)	(0.221)	(0.123)	(0.154)	(0.156)	(0.211)
belief_ diff_ gender					-0.063	-0.347*	-0.154*	0.237
					(0.089)	(0.188)	(0.082)	(0.176)
Constant	0.453	0.612	1.113*	0.269	0.426	0.373	0.951	0.187
	(0.429)	(0.420)	(0.541)	(1.053)	(0.445)	(0.475)	(0.557)	(0.914)
No. of observations	140	42	52	46	140	42	52	46

Notes: Standard errors in parentheses are clustered at the pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

ability are insignificant in both columns. Theory-of-mind ability is marginally associated with own best-fit narrative, while age is positively associated with it. Overall, Stage 1 accuracy is driven more by the informativeness of participant's data graph than by demographic or cognitive characteristics.

4.4.3 Consensus among conflicting-narrative pairs

We next examine predictors of consensus among participants who initially choose different narratives. Table 8 reports linear probability models for the full sample and separately by treatment. The dependent variable equals one if a participant in a conflicting-narrative pair chooses the same narrative as her partner in Stage 2. Columns (1)–(4) include individual traits and an indicator for whether the two matched participants have different own best-fit narratives, *data_ diff_ bestfit*. Columns (5)–(8) additionally include *belief_ diff_ gender*, an indicator for whether the participant believes her partner is of a different gender.[14]

Columns (1)–(4) and Columns (5)–(8) show similar significance patterns, so we focus on Columns (1)–(4). In the pooled specification, no variable significantly predicts consensus, consistent with the aggregate evidence that convergence is driven more by the treatment environment than by stable individual traits.

[14]Because beliefs about the partner's gender may be influenced by individual traits, we include this variable separately to avoid bad-control problems.

The treatment-specific estimates are more heterogeneous. Theory-of-mind ability is positively associated with consensus only in *Communication*. This is plausible because when participants lack direct access to their partner's data, unlike in *Share*, and have no prior reasoning task, unlike in *Reasoning*, the ability to infer another person's perspective may facilitate agreement.

The coefficients on *data_diff_bestfit* are insignificant in all specifications. Thus, conditioning on being in a conflicting-narrative pair, the fact that participants' personal data favor different narratives does not significantly predict consensus. This reinforces the evidence in Section 4.2.2, indicating that disagreement in personal best-fit narratives is not, by itself, a strong predictor for consensus failure.

Finally, the coefficients on *belief_diff_gender* are weakly negative in *Communication* and *Share*, suggesting that participants who believe their partner has a different gender are slightly less likely to reach consensus. This pattern does not appear in *Reasoning*, where choices rely more heavily on deliberative reasoning.

4.4.4 Stage 2 true-narrative correctness

Table 9 reports linear probability models for true-narrative correctness in Stage 2, for the full sample and by treatment. The dependent variable equals one if the participant chooses the true narrative after communication. Columns (1)–(4) include individual traits, the clarity of personal and pair data, and an indicator for Narrative 2 being true. Columns (5)–(8) additionally include variables potentially shaped by individual traits and data primitives, such as failing to choose the true narrative in Stage 1 (*wrongS1*), being in a conflicting-narrative pair (*conflicting*), and their interaction.

Columns (1)–(4) show limited effects of individual traits. Theory-of-mind ability is weakly positively associated with Stage 2 true-narrative correctness only in the *Reasoning*, while cognitive ability is negatively associated with correctness in the pooled and *Communication* specifications. Among data primitives, personal data clarity no longer predicts correctness after communication. Pair-level data clarity is weakly positive in the pooled and *Reasoning* specifications, suggesting higher correctness when the combined data more clearly favors one narrative, but this pattern does not appear in *Share*. Correctness remains significantly higher when Narrative 2 is true, consistent with stage 1 results. This effect disappears in Columns (5)–(8) after controlling for *wrongS1*, suggesting

Table 9: Determinants of choosing the true narrative in Stage 2

	(1) All	(2) Communication	(3) Share	(4) Reasoning	(5) All	(6) Communication	(7) Share	(8) Reasoning
Theory-of-Mind Ability	0.007	-0.003	0.012	0.016*	0.008*	0.007	0.006	0.013*
	(0.005)	(0.009)	(0.009)	(0.009)	(0.004)	(0.008)	(0.008)	(0.007)
Cognitive Ability	-0.055***	-0.089***	-0.041	-0.023	-0.047***	-0.070***	-0.069**	-0.004
	(0.019)	(0.033)	(0.034)	(0.037)	(0.017)	(0.024)	(0.031)	(0.036)
Male	0.046	-0.007	0.012	0.121	0.020	-0.051	0.019	0.098
	(0.057)	(0.115)	(0.095)	(0.103)	(0.048)	(0.083)	(0.089)	(0.076)
Age	0.001	0.003	-0.003	0.013	-0.005	-0.014	0.013	-0.008
	(0.014)	(0.024)	(0.019)	(0.031)	(0.011)	(0.018)	(0.016)	(0.021)
data_pair_SSEgap	1.031*	0.307	1.116	1.954*	0.201	-0.203	0.334	0.746
	(0.587)	(1.110)	(0.833)	(1.033)	(0.416)	(0.745)	(0.670)	(0.877)
data_own_SSEgap	-0.260	-0.070	-0.176	-0.372	-0.513	-0.452	-0.921	-0.528
	(0.568)	(0.908)	(0.874)	(1.005)	(0.315)	(0.444)	(0.725)	(0.846)
Narrative 2 is true	0.313***	0.275**	0.379***	0.262**	0.042	-0.051	0.099	-0.024
	(0.064)	(0.128)	(0.121)	(0.113)	(0.054)	(0.112)	(0.112)	(0.087)
wrongS1					-0.879***	-0.908***	-0.909***	-0.885***
					(0.050)	(0.094)	(0.105)	(0.090)
conflicting					-0.261***	-0.321**	-0.280***	-0.202*
					(0.059)	(0.121)	(0.092)	(0.111)
wrongS1×*conflicting*					0.737***	0.763***	0.885***	0.604***
					(0.076)	(0.131)	(0.122)	(0.162)
Constant	0.589**	0.816	0.552	0.043	1.141***	1.488***	0.872**	0.968**
	(0.297)	(0.504)	(0.414)	(0.632)	(0.219)	(0.359)	(0.349)	(0.463)
N	342	112	110	120	342	112	110	120

Notes: Standard errors in parentheses are clustered at the pair level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

that it operates largely through Stage 1 correctness.[15]

Columns (5)–(8) show that Stage 2 correctness is shaped more by prior choices and disagreement than by data clarity itself. The large negative coefficient on *wrongS1* reflects persistence in incorrect choices: participants who choose the wrong narrative in Stage 1 are less likely to choose the true narrative in Stage 2. More importantly, the interaction *wrongS1*×*conflicting* is large, positive, and significant in all treatments. This reproduces the pattern in Section 4.2.4, that is, participants who initially choose incorrectly are more likely to correct their choice when they learn that their partner initially chose a different narrative. Initial disagreement appears to trigger reconsideration.

Result 7 (Individual traits and data-environment primitives)**.** *Demographic and cognitive*

[15]Because the true narrative was randomly assigned at the matching-group level, Narrative 2 is not perfectly balanced across treatments: it is true for 31% of participants in *Communication*, 58% in *Share*, and 53% in *Reasoning*. Since Narrative 2 is associated with higher correctness, this imbalance may understate Stage 1 correctness in *Communication*. However, Stage 1 outcomes are similar across treatments: true-narrative correctness is 0.634, 0.691, and 0.675, and own best-fit correctness is 0.616, 0.645, and 0.708 in *Communication*, *Share*, and *Reasoning*, respectively. None of the pairwise differences is statistically significant. Thus, the imbalance is unlikely to drive Stage 2 treatment differences.

variables play at most a limited role in explaining narrative choices. By contrast, data-environment primitives are more predictive. Participants are more likely to choose the true or best-fit narrative when their own data more clearly distinguish between the two narratives, and when Narrative 2 is the true narrative.

In summary, we find that individual accuracy is driven primarily by the difficulty of the inference problem rather than by demographic or cognitive characteristics. Gender is not systematically related to narrative choice, and theory-of-mind and cognitive ability are not robust predictors of Stage 1 true-narrative correctness. The treatment effects documented above should therefore be interpreted as changes in how participants process and communicate evidence, rather than as consequences of demographic composition.

5 Conclusion

Disagreement over facts often persists even when people share a common objective and face a common underlying truth. This paper studies the extent to which free-form interpersonal communication can address such disagreement. In a controlled experiment on narratives about the sources of gender wage gaps, participants observe limited private data, choose between two competing narratives, communicate with a partner, and then choose again. This design allows us to distinguish two functions of communication that are often intertwined in practice, sharing private evidence and conveying how the evidence should be interpreted.

The results show that communication is effective at producing agreement, but agreement is not equivalent to truth-finding. Under free-form communication, initially conflicting pairs often converge on a common narrative, yet this convergence does not significantly increase the probability of choosing either the true underlying narrative or the narrative best supported by the available data. Direct data sharing helps participants use combined evidence more effectively, but does not ensure convergence to the true narrative. By contrast, prompting participants to articulate their reasoning increases the likelihood of choosing the best-fitting narrative, while reducing consensus among pairs who initially disagree. The central lesson is a trade-off, as communication protocols that discipline reasoning may also make narratives harder to revise.

These findings have broader implications for understanding real-world disagreement.

In many policy and social debates, disagreement persists not only because people hold different information, but also because they organize and interpret information through different mental models. Simply bringing people into conversation may generate superficial consensus without improving the quality of inference. Conversely, encouraging people to make their reasoning explicit may improve evidence-based judgment, but it can also make disagreement more persistent when private evidence supports different conclusions. For policy communication, deliberative forums, and online discussion design, the relevant question is therefore not only whether people communicate, but which dimension of communication improves, such as access to data, articulation of reasoning, or willingness to revise.

Several limitations suggest useful directions for future research. Our experiment studies a stylized binary narrative environment with aligned incentives, a common-value true state, and short-run communication between pairs. Real-world disagreements often involve richer narrative spaces, endogenous narrative construction, strategic motives, identity concerns, and repeated interactions. Future work could examine whether the same trade-off appears when narratives are not fixed in advance, when participants have expressive or reputational incentives, or when communication takes place in larger groups.

References

Aina, Chiara, "Tailored stories," *Unpublished manuscript*, 2023, *12*, 52.

_ **and Florian Schneider**, "Weighting competing models," 2025.

Ambuehl, Sandro and Heidi C Thysen, "Choice with Competing Models: An Experimental Study," 2025.

Baron-Cohen, Simon, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb, "The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism," *Journal of Child Psychology and Psychiatry*, 2001, *42* (2), 241–251.

Barron, Kai and Tilman Fries, "Narrative persuasion," 2025.

_ **, Heike Harmgart, Steffen Huck, Sebastian O Schneider, and Matthias Sutter**, "Discrimination, narratives, and family history: An experiment with jordanian host and syrian refugee children," *Review of Economics and Statistics*, 2023, *105* (4),

1008–1016.

Bénabou, Roland and Jean Tirole, "Identity, morals, and taboos: Beliefs as assets," *The Quarterly Journal of Economics*, 2011, *126* (2), 805–855.

Braghieri, Luca, Peter Schwardmann, and Egon Tripodi, "Talking across the aisle," Technical Report, Mimeo 2025.

Charles, Constantin and Chad Kendall, "Causal narratives," *Available at SSRN 4669371*, 2023.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

_ and _, "News media as suppliers of narratives (and information)," *arXiv preprint arXiv:2403.09155*, 2024.

_ , Simone Galperti, and Ran Spiegler, "False narratives and political mobilization," *Journal of the European Economic Association*, 2025, *23* (3), 983–1027.

Enke, Benjamin, "What You See Is All There Is," *The Quarterly Journal of Economics*, 2020, *135* (3), 1363–1398.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel, "Mental Models and Learning: The Case of Base-Rate Neglect," *American Economic Review*, March 2024, *114* (3), 752–82.

Fang, Ximeng, Sven Heuser, and Lasse S Stötzer, "How in-person conversations shape political polarization: Quasi-experimental evidence from a nationwide initiative," *Journal of Public Economics*, 2025, *242*, 105309.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental economics*, 2007, *10* (2), 171–178.

Fréchette, Guillaume R, Emanuel Vespa, and Sevgi Yuksel, "Extracting models from data sets: An experiment," *Available at SSRN 5026751*, 2024.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning Through Noticing: Theory and Evidence from a Field Experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Harrs, Sören, Lara Marie Müller, and Bettina Rockenbach, "How optimistic and pessimistic narratives about COVID-19 impact economic behavior," Technical Report, ECONtribute Discussion Paper 2021.

Iwasaki, Ichiro and Xinxin Ma, "Gender wage gap in China: a large meta-analysis,"

Journal for Labour Market Research, 2020, *54* (1), 17.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler, "Experimental tests of the endowment effect and the Coase theorem," *Journal of Political Economy*, 1990, *98* (6), 1325–1348.

Kendall, Chad and Ryan Oprea, "On the complexity of forming mental models," *Quantitative Economics*, 2024, *15* (1), 175–211.

Liu, Manwei and Sili Zhang, "Counteracting Narratives: Evidence from An Online Experiment," *The Economic Journal*, 2025, p. ueaf038.

Ma, Xinxin, "Internet use and gender wage gap: evidence from China," *Journal for Labour Market Research*, 2022, *56* (1), 15.

_ , "Gender role attitudes and the gender wage gap: evidence from China," *Journal of the Asia Pacific Economy*, 2025, *30* (2), 329–358.

Olea, José Luis Montiel, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat, "Competing models," *The Quarterly Journal of Economics*, 2022, *137* (4), 2419–2457.

Rabin, Matthew and Joel L. Schrag, "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics*, 1999, *114* (1), 37–82.

Roos, Michael and Matthias Reccius, "Narratives in economics," *Journal of Economic Surveys*, 2024, *38* (2), 303–341.

Santoro, Erik and David E Broockman, "The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments," *Science Advances*, 2022, *8* (25), eabn5515.

Schwardmann, Peter, Egon Tripodi, and Joël J. van der Weele, "Self-Persuasion: Evidence from Field Experiments at International Debating Competitions," *American Economic Review*, April 2022, *112* (4), 1118–46.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

_ **and** _ , "Sharing Models to Interpret Data," 2025.

Shiller, Robert J., "Narrative Economics," *American Economic Review*, April 2017, *107* (4), 967–1004.

Shiller, Robert J, "Narrative economics: How stories go viral and drive major economic events," 2020.

Staw, Barry M, "Knee-deep in the big muddy: A study of escalating commitment to a

chosen course of action," *Organizational Behavior and Human Performance*, 1976, *16* (1), 27–44.

Thomson, Keela S and Daniel M Oppenheimer, "Investigating an alternate form of the cognitive reflection test," *Judgment and Decision making*, 2016, *11* (1), 99–113.

Wang, Qianqian, Tsun-Feng Chiang, and Jing Jian Xiao, "Attitude toward gender inequality in China," *Humanities and Social Sciences Communications*, 2024, *11* (1), 1–14.

Yang, Yang and Jill E Hobbs, "The power of stories: Narratives and information framing effects in science communication," *American Journal of Agricultural Economics*, 2020, *102* (4), 1271–1296.

Zhang, Cheng, Xuan Tian, Xiaozhong Yang, Bing Xu, and Qinghai Li, "The iron-out effect of digital economy: A discussion on gender wage rate discrimination for working hours," *Journal of Business Research*, 2023, *156*, 113399.

Appendix A Data generating process

This appendix describes how we construct the two data pools used in the experiment. Each pool corresponds to one possible true narrative about the gender wage gap. The calibration follows the Oaxaca–Blinder decomposition estimates reported by Zhang et al. (2023) and Ma (2022).

The Oaxaca–Blinder decomposition separates the observed average wage gap into an *explained* component, due to group differences in productivity-related characteristics, and an *unexplained* component, often interpreted as the residual gap that remains after conditioning on those characteristics. Let $G_i \in \{0, 1\}$ indicate whether worker i is male, let X_i denote productivity, and let Y_i denote wage. In our experimental setting, the the data generating process is given by[16]

$$Y_i = \alpha + \beta X_i + \gamma G_i + \varepsilon_i, \tag{A.1}$$

where ε_i is an idiosyncratic noise term with mean zero and is orthogonal to G_i and X_i. Let $\mu_M = \mathbb{E}[X_i|G_i = 1]$ and $\mu_F = \mathbb{E}[X_i|G_i = 0]$ be the mean productivities of male and female workers, respectively. Then, by (A.1), the gender wage gap can be decomposed as

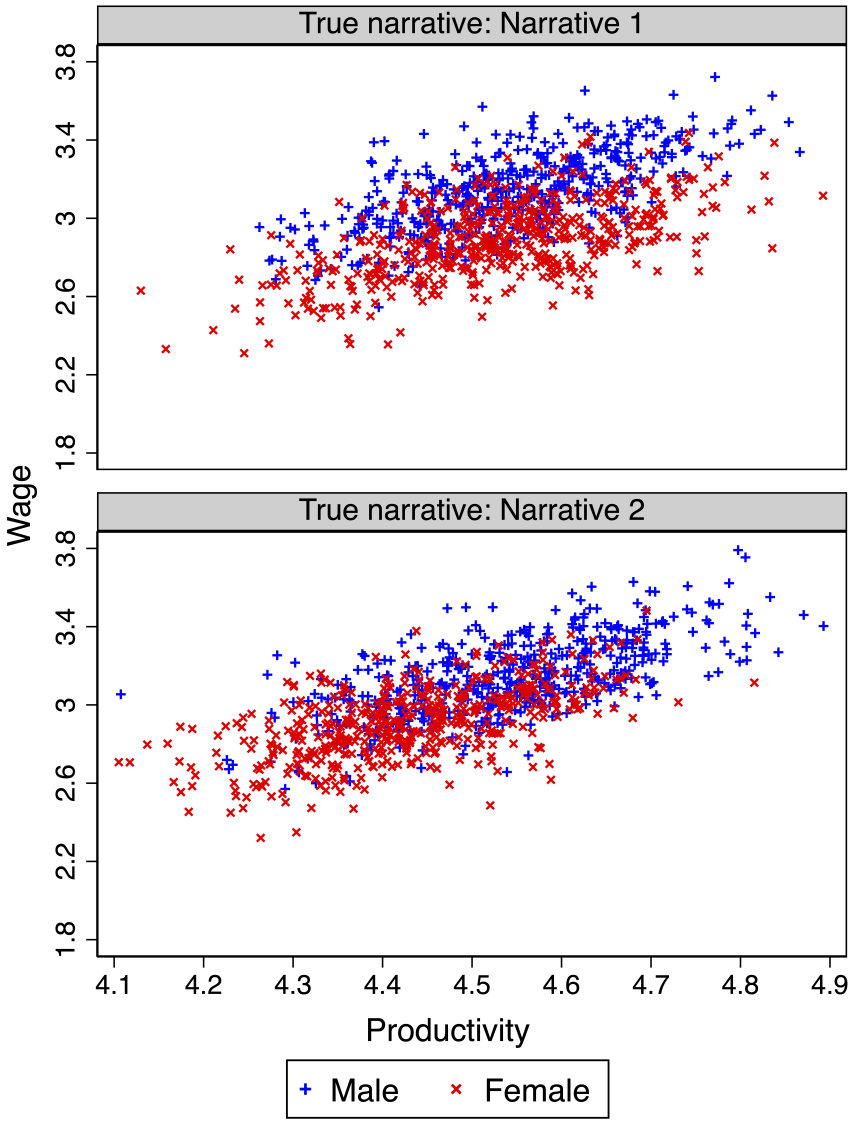
$$\Delta Y = \mathbb{E}[Y_i \mid G_i = 1] - \mathbb{E}[Y_i \mid G_i = 0] = \underbrace{\beta(\mu_M - \mu_F)}_{\text{explained}} + \underbrace{\gamma}_{\text{unexplained}}. \tag{A.2}$$

The explained component comes from gender differences in productivity, while the unexplained component is the same-productivity male wage premium. By (A.2), the share of unexplained component in observed gender wage gap thus equals $u = \gamma/\Delta Y$.

Following Zhang et al. (2023), we fix the average male productivity at $\mu_M = 4.544$, and set the average wages to $\bar{Y}_M = 3.144$ for males and $\bar{Y}_F = 2.892$ for females. The gender wage gap is then fixed at $\Delta Y = 0.252$. The two narratives differ only in the share of this gap attributed to the unexplained component. Narrative 1 uses $u_1 = 0.9472$ from Zhang et al. (2023); Narrative 2 uses $u_2 = 0.512$ from Ma (2022). Therefore, under each narrative $k \in \{1, 2\}$, the implied male premium is $\gamma_k = u_k \Delta Y$ and the average gender productivity gap is $\mu_M - \mu_{F,k} = (1 - u_k)\Delta Y$. Finally, the intercept coefficient in (A.1) is set to match the average male wage; that is, $Y_M = \alpha_k + \delta_k + \mu_k$ for each narrative k. Equivalently, this implies $\alpha_k = \bar{Y}_M - \mu_M - \gamma_k = -1.4 - \gamma_k$ for each narrative $k \in \{1, 2\}$.

[16] In the empirical framework of Zhang et al. (2023) and Ma (2022), the dependent variable Y_i is the logarithm of wage, and they allow for multiple observable productivity-related characteristics. To simplify our experiment, we directly interpret Y_i as wage and consider only a single productivity index X_i.

Figure A.1: Datasets



For each narrative $k \in \{1, 2\}$, we simulate a pool of 1,000 independent observations. For each observation i, gender is drawn as $G_i \sim \text{Bernoulli}(1/2)$, productivity is drawn from $X_i \mid G_i = 1 \sim N(\mu_M, 0.12^2)$ for male workers and $X_i \mid G_i = 0 \sim N(\mu_{F,k}, 0.12^2)$ for female workers. Finally, we set $\beta = 1$ and generate each idiosyncratic error term by $\varepsilon_i \in N(0, 0.15^2)$. Then, by (A.1), each worker i's wage is $Y_i = \alpha_k + X_i + \gamma_k G_i + \varepsilon_i$. This construction gives the same average gender wage gap, 0.252, under both narratives.

The two resulting data pools are shown in Figure A.1. The horizontal axis in the graph is productivity X_i, and the vertical axis is wage Y_i. In the experiment, when Narrative k is selected as the true narrative, each participant's personal data graph is generated by sampling 20 observations from the corresponding pool.

Appendix B Experimental instructions

Please note that the original instruction was in Chinese, and what follows is an English translation. Instructions that apply to all treatments are shown in regular font. Paragraphs and sections specific to each treatment are presented in italics, with treatment names shown in parentheses.

B.1 Welcome

Welcome to this decision-making experiment! Please read the following instructions carefully. During the experiment, please stay quiet and avoid talking to others. If you have any questions, raise your hand, and an experimenter will assist you privately. This experiment will take approximately 60 minutes.

You will participate in the experiment with other participants, each seated at a separate computer. You will not be able to know others' identities. This experiment is anonymous. Neither the experimenters nor the other participants can link your decisions to your desk number or your identity.

Your earnings in the experiment will be calculated in Chinese *yuan*. In addition, you will receive a show-up fee of 20 *yuan* for participating, which will be added to your total earnings. Your payment will be given privately at the end of the experiment.

(Communication & *Share*) *The experiment consists of two stages. First, you will carefully read the instructions for the first stage. After reading, you will need to correctly answer a set of questions to ensure you understand the experiment and then make your decisions. Once you complete your decisions for the first stage, you will proceed to read the instructions for the second stage, answer another set of understanding questions, and then make your decisions.* ***At the end of the experiment, the computer will randomly select one of the two stages with equal probability to determine your total earnings. The earnings from the non-selected stage will not be included in your total earnings.***

(Reasoning) *The experiment consists of three stages. In each stage, you will first carefully read the instructions for that stage. After reading, you will need to correctly answer a set of questions to ensure you understand the experiment and then make you decisions. Your earnings in the experiment will be distributed in two parts:* ***(1) Earnings***

from the first and third stages will be paid immediately after the experiment. At the end of the experiment, the computer will randomly select either the first or the third stage with equal probability to determine your total earnings. The earnings from the non-selected stage will not be included in your total earnings. (2) Earnings from the second stage will be paid through Ancademy 2 to 4 weeks after the experiment. A detailed explanation about this process will be provided in the second-stage instructions.

B.2 Stage 1: Inferring the causes of the gender wage gap

B.2.1 Experimental background

Statistical data shows that, in today's labor markets, the average wage of women is generally lower than that of men. For example, in 2023, the average monthly salary of male workers was 9942 yuan, while the average monthly salary of female workers was 8689 yuan, which is 1253 yuan, or 12.6%, lower than that of male workers.

What causes the average wage of men to be higher than that of women? Research indicates that, in Chinese society, the gender wage gap can generally be attributed to two main factors. The first is productivity differences, meaning that men tend to have higher productivity levels than women. The second includes factors that cannot be explained by productivity differences. For example, even when a male worker and a female worker have the same level of productivity, the male worker may still earn higher wages than the female worker. More specifically, the gender wage gap can be divided into two components based on its sources:

(1) **Factors of productivity differences:** The productivity level, also known as labor endowment, is a comprehensive indicator that measures all observable characteristics related to an individual's productivity, including but not limited to education, work experience, working hours, family background, marital and childbearing status, age, political affiliation, and household registration. Generally, individuals with higher productivity tend to earn higher wages. For example, individuals with higher education typically receive higher wages, and similarly, those with more work experience also tend to earn higher wages. Therefore, if, on average, men have higher productivity levels than women, the average wage of men will be higher than that

of women. In this case, if productivity differences between men and women indeed exist, these factors can partly or fully explain the gender wage gap. We refer to the portion of the gender wage gap that can be explained by productivity differences as the **"explained component"**.

(2) **Factors unexplained by productivity differences:** Generally, productivity differences alone cannot fully explain the gender wage gap. In addition to the "explained component," there remains a portion of the gender wage gap that cannot be explained by productivity differences. Specifically, even when a man and a woman have the same level of productivity, the man may still earn a higher wage than the woman due to other factors. We refer to the portion of the gender wage gap that cannot be explained by observable productivity differences as the **"unexplained component"**.

In summary, the gender wage gap = "explained component" + "unexplained component".

In reality, what proportion of the gender wage gap can be attributed to the "explained component" (i.e., observable productivity differences between the genders) and what proportion can be attributed to the "unexplained component" (i.e., even when productivity levels are the same, men still earn higher wages than women)?

There are numerous academic studies that address this question. In today's experiment, we select two representative studies, referred to as **Study A** and **Study B**. Both studies analyze real wage data of male and female workers in China from recent years, using similar research methods to draw conclusion about the proportion of the "explained component". However, due to differences in the data collection and analysis approaches, Study A and Study B reach different conclusions, as outlined below:

- **Study A finds that less than 10% of the gender wage gap is attributed to the "explained component", while more than 90% is attributed to the "unexplained component".**

The conclusion of Study A reveals that, in practice, the observable productivity difference between men and women are minimal, indicating that the gender wage gap cannot be attributed to productivity differences. In other words, even though men and women have nearly identical productivity levels on average, men still earn higher wages

than women. Therefore, the primary cause of the gender wage gap is the "unexplained component".

- **Study B finds that "explained component" and "unexplained component" each account for approximately 50% of the gender wage gap.**

The conclusion of Study B reveals that, in practice, men's average productivity level is significantly higher than that of women, which partly contributes to the higher average wages of men compared to women. At the same time, there also exist the unexplained component, meaning that, even with the same level of productivity, men still earn higher wages than women on average. This further contributes to men's higher average wages. Therefore, both the "explained component" and the "unexplained component" together cause the gender wage gap, each accounting for approximately 50%.

B.2.2 Experimental procedure

In today's experiment, you will see data of 20 workers, including their productivity levels and wages. **You will need to infer the proportion of the "explained component" and the "unexplained component" in the gender wage gap based on the data you observe from these 20 workers.**

In today's experimental setting, all data you observe are generated by computer based on the conclusions from either Study A or Study B. The detailed data generation process is as follows:

- Before the experiment begins, the computer program will randomly select either Study A or Study B, with each having a 50% probability of being selected. The computer program will then set the proportion of the "explained component" and the "unexplained component" according to the selected study. Specifically, for the data generated by the computer, if Study A is selected, the "explained component" will account for less than 10%; if Study B is selected, the "explained component" will account for approximately 50%.

- Next, the computer program will generate data for 20 workers based on the proportions of the "explained component" and the "unexplained component". The data will include each worker's gender, productivity level, and wage. Specifically:

(1) Each worker's gender is randomly and independently assigned by the computer program, with a 50% probability of being male or female.

(2) Each worker's productivity level and wage are also generated by the computer program, with the generation process being independent across workers. For both male and female workers, there is a positive relationship between wage and productivity level: as the productivity level increases, the average wage for that productivity level increases accordingly, and each worker's wage will fluctuate around the average. Additionally, for every unit increase in the productivity level, the increase in the average wage is the same for both male and female workers.

(3) Given the same productivity level, the wage gap between male and female workers depends on the proportion of the "unexplained component" selected. For any given productivity level, if the proportion of the "unexplained component" is higher, male workers' average wage will be higher compared to that of female workers; conversely, if the proportion of the "unexplained component" is lower, average wages of male and female workers will be closer.

(4) The productivity differences between male and female workers depend on the proportion of the "explained component." If the proportion of the "explained component" is higher, male workers' average productivity level will be higher than that of female workers. Conversely, if the proportion of the "explained component" is lower, the difference in the average productivity levels between two genders will be smaller, and the difference in the average wage between two genders at the same productivity level will be larger.

Note that the sum of the proportions of the "explained component" and the "unexplained component" must equal 100%. Therefore, if the proportion of the "explained component" increases, then the proportion of the "unexplained component" must decrease, and vice versa.

The following graph shows an example of the data you will observe, where each points represents a worker. Female workers are indicated by red points and male workers are indicated by blue points. The horizontal coordinate represents each worker's productivity level, and the vertical coordinate represents their wage. The wage is expressed as the

logarithmic (log) value of the worker's hourly wage. Although the units of the productivity has no specific meaning, a higher number indicates a higher productivity level. The average wages for male and female workers in this example of 20 data points are displayed in parentheses on the right side of the graph. For example, in the example graph, the numbers in parentheses show that the average wage for male workers is 3.058, while the average wage for female workers is 2.732, indicating that male workers indeed have a higher average wage than female workers.

In the experiment, your task is to observe your 20 data points and infer the proportions of the "explained component" and the "unexplained component" that you think are reflected in the data. Based on this, you will determine whether the data aligns more closely with Study A or Study B. Note that since the data is generated according to either Study A or Study B, the proportions you infer must be consistent with conclusions of one of two studies.

Your earnings in this stage depend on whether you correctly infer the proportions of the "explained component" and the "unexplained component", or, equally, whether you correctly choose the conclusions of Study A or Study B. If your choice is correct, you will earn 40 *yuan*; if your choice is wrong, you will earn zero.

B.2.3 Control questions of Stage 1

(i) If the proportion of the "explained component" is 10%, what is the proportion of the "unexplained component"? (a) 10% ; (b) 50% ; (c) 90% ; (d) Cannot be determined;

(ii) If the proportion of the "unexplained component" increases, for male and female workers having the same productivity level, the difference in their average wages will be: (a) larger ; (b) smaller ; (c) the same ; (d) cannot be determined;

(ii) Given the gender wage gap, if the difference in the average productivity levels between two genders increases, the proportion of the "unexplained component" will be: (a) larger ; (b) smaller ; (c) the same ; (d) cannot be determined.

B.3 (Reasoning) *Stage 2: Providing advice for future experiment participants*

In this stage, your task is to write advice on how to make decisions in Stage 1 for participants who will participate in a future experiment. Specifically, you need to help a future participant decide between Study A and Study B based on their data graph.

The procedure of the future experiment is as follows: The experimenter will recruit new experimental participants at another university within the next two to three weeks. They will face the same decision as you did in Stage 1, where they will observe 20 data points and decide between Study A and Study B. ***However, before making their decision, each future participant will see three pieces of advice written by participants in the current experiment and will be asked to choose the one they find "most helpful".*** *Specifically, a future participant, who is randomly selected by the computer, will see your advice along with the advice from two other participants in the current experiment (making a total of three pieces of advice). The order in which the advice is presented will be random. All advice is anonymous, so the future participants will not know the identity of the advice providers.*

Note that for the future participants who will see your advice, the study selected by the computer will be the same as yours (i.e., either Study A or Study B for both of you). However, the 20 data points they will observe are randomly generated by the computer based on the selected study, so it is highly likely that their data will differ from yours.

If your advice is selected as the "most helpful", you will earn 20 yuan; if it is not selected, you will earn zero. Your advice will be seen by only one future participant, so you will either earn 20 yuan or zero for this stage. Your earnings for this stage will be transferred to your Ancademy account within two to four weeks after the future participants have completed their experiment. Note that Ancademy will only use your student ID to distribute earnings. The platform and experimenters will not use or disclose your personal information to any third parties.

B.3.1 Control questions of Stage 2

(i) Which of the following statements is correct? The studies selected by the computer for the future participant and me: (a) must be the same; (b) must be different; (c) might be the same or different;

(ii) Which of the following statements is correct? (a) The 20 data points observed by the future participant are exactly the same as mine; (b) The 20 data points observed by the future participant and me are generated independently by the computer.

B.4 Stage 3: Random matching and free discussion

In the previous stage, you and the other participants in the room have followed the same experimental process and inferred the proportions of the "explained component" and the "unexplained component". **In this stage, you will be randomly matched with one of the other participants in the room.** The matching is anonymous, and neither you nor your matched partner will be able to identify each other.

In this stage, you and your partner can freely discuss the proportions of the "explained component" and the "unexplained component". The discussion will last for a maximum of **15 minutes**. If both of you feel that the discussion is sufficient, you can choose to end it earlier. During the discussion, do not mention any personal identity information or use inappropriate language.

(Share) ***During the discussion, both of you and your partner will be able to view the 20 data points each of you observed in the previous stage.***

After the discussion, you will need to determine whether the data aligns with the conclusions of Study A or Study B again, which involves inferring the proportions of the "explained component" and the "unexplained component".

Note that the study selected by the computer, either Study A or Study B, will remain the same throughout all stages of the experiment, and will be the same for both of you and your matched partner. However, since each participant's 20 data points observed in the experiment are randomly generated by the computer based on the selected study, and the data generation process is independent across participants, **the 20 data points you and your partner have observed may differ**.

(Share) ***During the free discussion, you will view three data graphs: one displaying the 20 data points you observed in the previous stage; another displaying the 20 data points your partner observed in the previous stage; and a third is a combination of both you and your partner's data, consisting of 40 data points. In addition, you will also be informed of the study choices made by you and your partner in the previous stage. Your partner will also***

see the same information as you.

(Communication & Reasoning) ***During the free discussion, you will be able to review the data you observed in the previous stage. In addition, you will also be informed of the study choices made by you and your partner in the previous stage.***

The remaining time for the free discussion will be displayed in the upper right corner of the screen. If both of you wish to end the discussion earlier, you and your partner must click the "End Early" button. Only when both of you click this button will the free discussion end early. If at least one of you does not click the "End Early" button, the discussion will continue and automatically end after 15 minutes.

Your earning in this stage will be determined in the same way as in Stage 1. If your choice of Study A or Study B is correct, you will earn 40 *yuan*; if your choice is wrong, you will earn zero.

B.4.1 Control questions of Stage 2

(i) Which of the following statements is correct? For me and my matched partner, the study selected by the computer (either Study A or Study B): (a) might be the same; (b) must be the same; (c) must be different;

(ii) Which of the following statements is correct? The study selected by the computer (either Study A or Study B) in Stage 2: (a) must be the same as in Stage 1; (b) must be different from in Stage 1; (c) might be the same as in Stage 1;

(iii) Which of the following statements is correct? (a) The 20 data points observed by my partner in Stage 1 are exactly the same as mine; (b) The 20 data points observed by my partner and me in Stage 1 are generated independently by the computer.

B.5 Experimental screenshots

Figure B.1: Screenshot of the first-stage decision

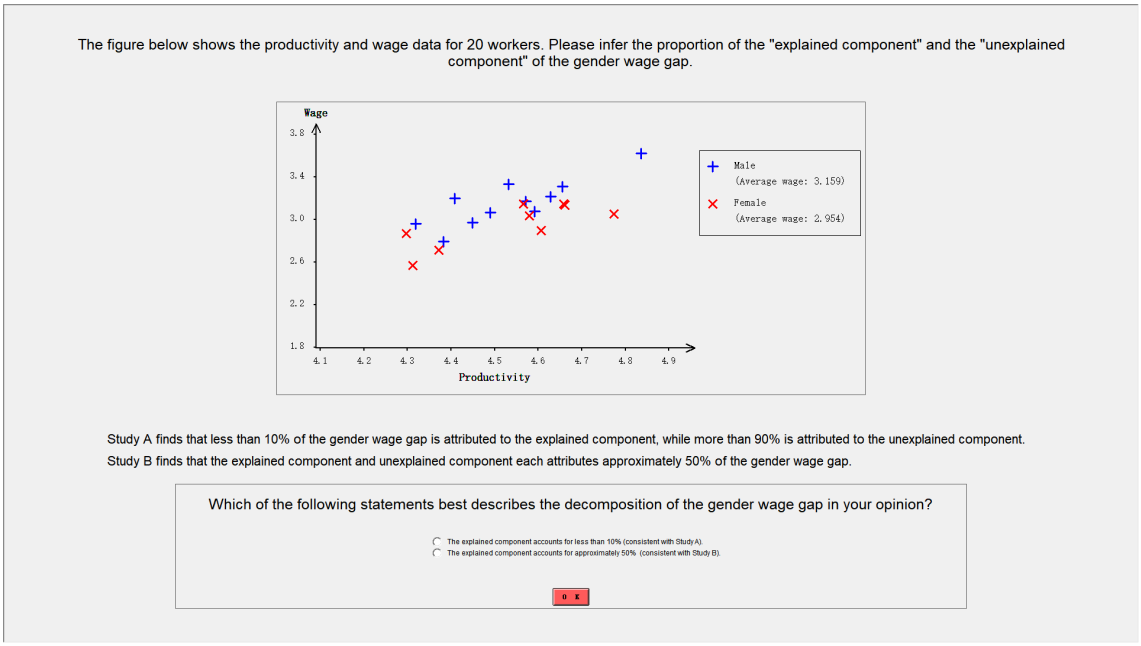


Figure B.2: Screenshot of the second-stage communication in the *Communication* (and *Reasoning*) treatment

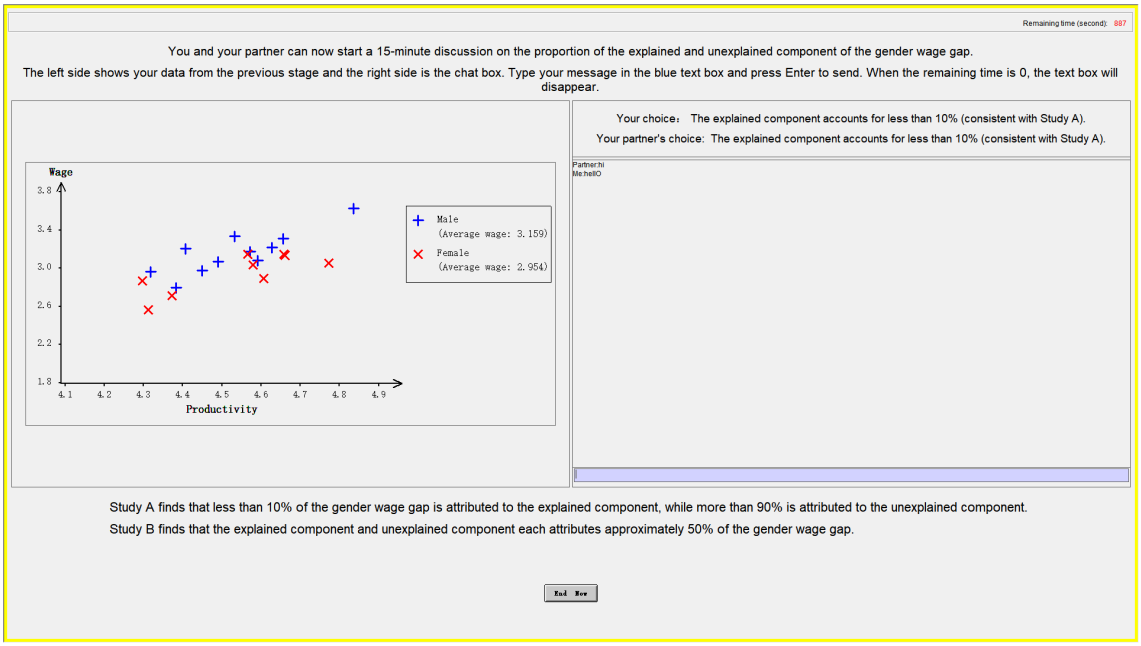


Figure B.3: Screenshot of the second-stage communication in the *Share* treatment

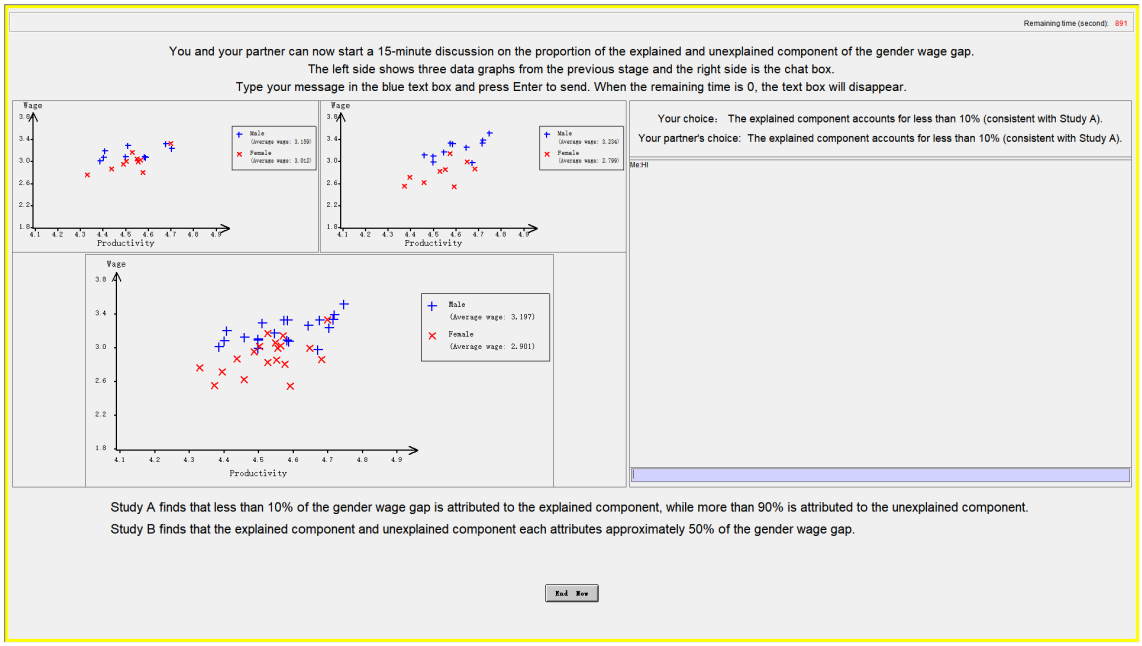
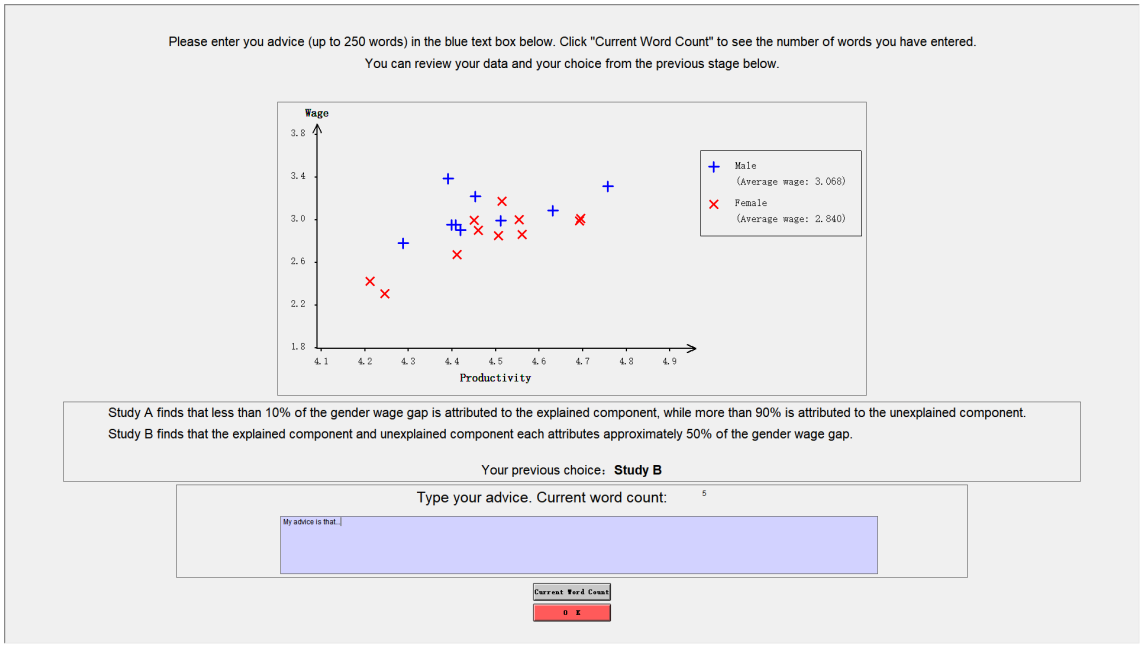


Figure B.4: Screenshot of the advice-writing stage in the *Reasoning* treatment

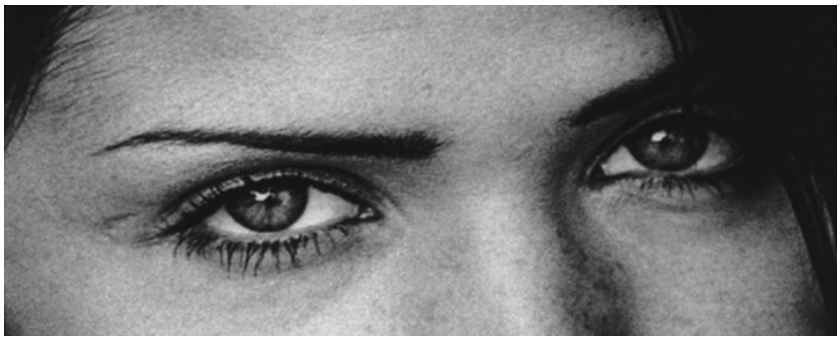


B.6 Elicitation of theory of mind and cognitive ability

In this section, we explain how we elicit the theory of mind and cognitive ability in the experiment. Note that the material below is not limited to the instructions shown directly to participants, although it includes those instructions.

To elicit theory of mind (ToM), we use the adult version of the *Reading the Mind in the Eyes Test* developed by Baron-Cohen et al. (2001). The original test contains 36 photographs of eyes; due to time constraints, we administered 18 photographs corresponding to the first half of the original test. For each photograph, participants selected one of four words that best described the depicted mental state. One practice question is provided before the test began. The test is limited to six minutes, and participants earned 0.5 CNY for each correct answer. Figure B.5 presents an example of the theory-of-mind test.

Figure B.5: A theory-of-mind test in the experiment (Note: The four response options are “joking", “flustered", “desire", and “convinced".)



We measure cognitive ability using two complementary tasks. First, participants complete three items from the Advanced Raven's Progressive Matrices, which measure abstract reasoning ability. Second, they answer three questions from the Cognitive Reflection Test of Thomson and Oppenheimer (2016), which measures the tendency to override intuitive but incorrect responses. The cognitive ability test is limited to six minutes, and participants earned 1 CNY for each correct answer. The Cognitive Reflection Test is presented below.

1. In a race, if you overtake the runner in the 2nd position, what position are you at?: (a) 1st; (b) 2nd; (c) 3rd; (d) 4th;

2. A farmer had 15 sheep, and all but 8 died. How many are left?: (a) 0; (b) 7; (c) 8; (d) 10;

3. How many cubic meters of dirt are there in a hole that is 3-meter deep, 3-meter wide, and 3-meter long?: (a) 0; (b) 3; (c) 9; (d) 27;

Appendix C Descriptions and summaries of variables

C.1 Descriptions of coding variables

Three research assistants independently coded the chat records using the rules summarized in Table C.1. For each pair of participants, coders first read the entire conversation and then coded each participant separately. They were instructed that only *Describing Method* allowed multiple selections, since participants could mention multiple methods; all other variables required a single selection.

The coding variables used in the analysis, derived from Table C.1, are presented in Table C.2. *Agreeing Data* and *Stating Consensus* were initially coded in more detailed categories, but were consolidated into the variables *diff_data* and *state_consensus*, respectively, because some original categories contained fewer than 10 observations.

Table C.1: Variables coded by coders

Variables	Coding rule
Confirming the Goal	1 = if mentioning that they and their partner have the same goal, i.e., they have the same "correct answer"; 0 = if neither of them.
Feeling Uncertain	1 = if mentioning uncertainty or hesitation about one's own choice at the **beginning of the conversation**; 0 = if neither of them.
Agreeing Data	1 = if believing their data is not similar to the partner's and agrees more with their own data; 2 = if believing their data is not similar to the partner's and agrees more with the partner's data; 3 = if believing their data is not similar to the partner's under other circumstances (e.g., being neutral, uncertain, or no relevant expression regarding their data); 0 = if **not** believing their data is not similar to the partner's (e.g., believing their data is similar to the partner's or no relevant expression).
Describing Data	1 = if describing their own data or graph to the partner; 0 = if **not** describing their own data or graph to the partner.
Describing Method	1 = if describing their own method or reasoning as "calculating (or comparing) the **productivity** or **wage levels** of at least one gender"; **2** = if describing their own method or reasoning as "discussing (or comparing) the **slopes** or **shapes** of the curve of at least one gender"; **3** = if describing their own method or reasoning as "comparing wage levels given the same productivity levels between two genders."; **4** = if describing any other method or reasoning; 0 = if **not** describing their method or reasoning to the partner.
Asking Data or Method	1 = if asking the partner for their data or graph; 2 = if asking the partner for their method or reasoning; 3 = if asking the partner for both their data or graph and their method or reasoning; 0 = if none of them.
Mentioning the Reality	1 = if mentioning unexplained factors in reality (e.g., discrimination); 2 = if mentioning explained factors in reality (e.g., gender differences in productivity); 3 = if mentioning both unexplained and explained factors in reality; 0 = if none of them.
Persuasion	1 = if **trying to persuade** the partner (e.g., stating they are more reasonable); 2 = if stating **being persuaded** by the partner (e.g., saying they will change their choice); 0 = if neither of them;
Stating Consensus	1 = if both participants **explicitly** express they have reached a consensus; 2 = if both participants **vaguely** express they have reached a consensus; 0 = if neither of them.

Notes: Only *Describing Method* allows multiple selections. All other variables require a single selection.

Table C.2: Coding variables

Variables	Description
same_goal	Equivalent to *Confirming the Goal*
uncertain	Equivalent to *Feeling Uncertain*
agree_owndata	Equivalent to *Agreeing Data*
diff_data	1= if *Agreeing Data* equals 1, 2, or 3 (i.e., believing their data is not similar to the partner's); 0 = if *Agreeing Data* equals 0.
describe_data	Equivalent to *Describing Data*
describe_method	1 = if *Describing Method* is greater than 0 (i.e., describing any own method or reasoning); 0 = if otherwise.
describe_method1	1 = if *Describing Method* equals 1 (i.e., describing their method or reasoning as "calculating (or comparing) the productivity or wage levels of at least one gender"); 0 = if otherwise.
describe_method2	1 = if *Describing Method* equals 2 (i.e., describing their method or reasoning as "discussing (or comparing) the slopes or shapes of the curve of at least one gender"); 0 = if otherwise.
describe_method3	1 = if *Describing Method* equals 3 (i.e., describing their method or reasoning as "comparing wage levels given the same productivity levels between two genders."); 0 = if otherwise.
describe_method4	1 = if *Describing Method* equals 4 or (i.e., if describing any other method or reasoning); 0 = if otherwise.
ask_data	1 = if *Asking Data or Method* equals 1 or 3 (i.e., asking the partner for their data or graph); 0 = if otherwise.
ask_method	1 = if *Asking Data or Method* equals 2 or 3 (i.e., asking the partner for their method or reasoning); 0 = if otherwise.
real_unexplained	1 = if *Mentioning the Reality* equals 1 (i.e., mentioning unexplained factors in reality); 0 = if otherwise.
real_explained	1 = if *Mentioning the Reality* equals 2 (i.e., mentioning explained factors in reality); 0 = if otherwise.
persuading	1= if *Persuasion* equals 1 (i.e., trying to persuade the partner); 0 = if otherwise.
being_persuaded	1= if *Persuasion* equals 2 (i.e., stating being persuaded by the partner); 0 = if otherwise.
state_consensus	1= if *Stating Consensus* equals 1 or 2 (i.e., both participants state they have reached a consensus); 0 = if otherwise.

C.2 Summary of individual characteristics and data-environment variables

Table C.3 presents the definitions and summary statistics for each participant's demographic, cognitive, and data-environment variables by treatment. Demographic variables, derived from the post-experiment survey, include each participant's gender, age, beliefs about their partner's gender (*belief_diff_gender* and *belief_same_gender*), and whether the two matched participants have the same gender (*same_gender*). Cognitive variables assess theory-of-mind abilities and cognitive abilities. Data-environment variables include an indicator for whether Narrative 2 is the true narrative, as well as measures of the difficulty in identifying the best-fit narrative based on personal data (*data_own_SSEgap*) and on pair data (*data_pair_SSEgap*). In addition, we construct an indicator for whether personal best-fit narratives of two matched participants differ (*data_diff_bestfit*). See Appendix D.1 for detailed constructions of the data-environment variables.

Table C.3: Summary of demographic, cognitive, and data-environment variables

Variables	Mean				Kruskal-Wallis tests (p-value)	Definition
	All	*Communication*	*Share*	*Reasoning*		
Theory-of-Mind Ability	9.406 (4.773)	9.580 (4.709)	9.236 (5.036)	9.400 (4.630)	0.832	The number of correct answers in the ToM test (minimum = 1, maximum = 18)
Cognitive Ability	3.991 (1.147)	3.777 (1.206)	4.155 (1.085)	4.042 (1.126)	0.071*	The number of correct answers in the cognitive ability test (minimum = 1, maximum = 6)
Male	0.216 (0.412)	0.179 (0.385)	0.291 (0.456)	0.183 (0.389)	0.071*	1 = male; 0 = female.
Age	20.6 (1.9)	20.7 (2.0)	20.4 (1.7)	20.5 (1.7)	0.277	Age in years (minimum = 17, maximum = 27)
belief_diff_gender	0.360 (0.481)	0.339 (0.476)	0.382 (0.488)	0.358 (0.482)	0.804	1 = if guessing the partner has a different gender; 0 = if not.
belief_same_gender	0.342 (0.475)	0.330 (0.472)	0.300 (0.460)	0.392 (0.490)	0.327	1 = if guessing the partner has the same gender; 0 = if not.
same_gender	0.673 (0.470)	0.714 (0.454)	0.600 (0.492)	0.700 (0.460)	0.141	1 = if the partner has the same gender; 0 = if not.
Narrative 2 is true	0.474 (0.500)	0.304 (0.462)	0.582 (0.496)	0.533 (0.501)	0.001***	1 = if Narrative 2 is the true narrative; 0 = if not.
data_own_SSEgap	0.051 (0.045)	0.057 (0.058)	0.050 (0.039)	0.047 (0.036)	0.668	The absolute difference between the SSE values of two narratives for the personal data. A higher value indicates the best-fit narrative is easier to identify. (minimum = 0.000, maximum = 0.401)
data_pair_SSEgap	0.090 (0.061)	0.095 (0.064)	0.088 (0.065)	0.088 (0.054)	0.720	The absolute difference between the SSE values of two narratives for the pair data. (minimum = 0.002, maximum = 0.243)
data_diff_bestfit	0.357 (0.480)	0.393 (0.491)	0.382 (0.488)	0.300 (0.460)	0.271	1 = if the best-fit narratives of the two matched participants' personal data differ; 0 = if they are the same.

Notes: Standard deviations are shown in parentheses. For Kruskal-Wallis tests, $n = 342$, $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We highlight two key results from Table C.3. First, 67% of all participants are randomly assigned to pairs with a partner of the same gender in the experiment. When asked whether they had made a guess about their partner's gender, 36% of participants

guessed that their partner had a different gender, 34% guessed that their partner had the same gender, and the remaining 30% did not make a guess. Second, for 36% of all participants, the best-fit narratives based on their own personal datasets differ from that of their partners, while for the remaining pairs, both participants' data suggest the same best-fit narrative.

Appendix D Methods for data analyses and supplementary tables

D.1 Methods for determining the best-fit narrative

In this appendix we explain our methods for constructing the measures of narrative fitness under a given dataset D — either a participant's private 20-observation dataset or a pair's combined 40-observation dataset. For narrative $k \in \{1, 2\}$, consider linear regression

$$Y_\ell = \alpha_k + \beta_k X_\ell + \gamma_k G_\ell + \varepsilon_\ell, \qquad \ell \in D, \tag{D.1}$$

where Y_ℓ is worker ℓ's wage, X_ℓ is worker ℓ's productivity, G_ℓ indicates whether worker ℓ is male, and ε_ℓ is an idiosyncratic error term. Let $\overline{Y}_{D,M}$ and $\overline{Y}_{D,F}$ denote, respectively, the average wages of male and female workers in dataset D. Then $\Delta Y_D = \overline{Y}_{D,M} - \overline{Y}_{D,F}$ gives the average gender wage gap for workers in D. Recall that the shares of unexplained components in the gender wage gap are $u_1 = 0.9472$ for Narrative 1 and $u_2 = 0.512$ for Narrative 2. Then, according to the discussion in Appendix A, the implied male wage premium under narrative k – the gender dummy coefficient γ_k in (D.1) – must satisfy[17]

$$\gamma_k = u_k \Delta Y_D \ . \tag{D.2}$$

For each narrative $k \in \{1, 2\}$, we compute the least-square estimates α_k and β_k of (D.1) under parametric restriction (D.2), and then obtain the resulting sum of squared residuals, denoted by $SSE_k(D)$.[18] Given dataset D, Narrative 1 is identified as the best-fit narrative if $SSE_1(D) < SSE_2(D)$; otherwise, Narrative 2 is the best-fit narrative.[19] Intuitively, the

[17]Note that we use the actual proportions of the unexplained component – 94.72% from Zhang et al. (2023) and 51.2% from Ma (2025) – in the estimation. The main results are robust to using the rounded proportions (90% and 50%) described in the experimental instructions.

[18]This is done by constructing $\tilde{Y}_\ell := Y_\ell - u_k \Delta Y_D G_\ell$ for each observation $\ell \in D$, and then conducting OLS estimation to $\tilde{Y}_\ell = \alpha_k + \beta_k X_\ell + \varepsilon_\ell$. The resulting sum of squared residuals is our object.

[19]In the unlikely event $SSE_1(D) = SSE_2(D)$, it is innocuous to identify either narrative as best-fit.

best-fit narrative is the one whose implied decomposition better explains the observed wage-productivity pattern from dataset D.

Finally, given each participant i and her private dataset D_i, we construct a natural measure $data\_own\_SSEgap_i$ to reflect the statistical power of dataset D_i in distinguishing the two candidate narratives. Specifically,

$$data\_own\_SSEgap_i = |SSE_1(D_i) - SSE_2(D_i)| \ . \tag{D.3}$$

A higher $data\_own\_SSEgap_i$ thus indicates a greater difference in fitness between the two narratives given i's private dataset D_i. As a result, the underlying personal data more clearly favor one narrative over the other, making it easier to identify the true underlying narrative. We also construct an analogous measure for the combined dataset of a matched pair. Let D_{ij} denote the combined dataset of participant i with partner j. Then

$$data\_pair\_SSEgap_{ij} = |SSE_1(D_{ij}) - SSE_2(D_{ij})| \ . \tag{D.4}$$

D.2 Methods for analyzing the drivers of treatment differences

We examine each coding variable, denoted as X, as a potential explanatory variable. The advantage of our approach is that it examines each potential explanatory variable individually, avoiding the issues associated with variable selection in multiple regression models. Specifically, the methodology proceeds in two steps. For each pair of treatments and for each coding variable X, the first step involves conducting the following regression using the linear probability model:

$$X = \alpha_1 + \beta_1 TreatDummy + \varepsilon_1, \tag{D.5}$$

where $TreatDummy$ is a dummy variable equal to 1 for the treatment selected as the benchmark and 0 for the other treatment. The coefficient β_1 represents the *treatment effect*, as reported in Column (1) of Tables D.1, D.2, and D.3. A significant β_1 indicates that X differs significantly between the two treatments.

The second step varies depending on the key outcomes on which we focus. Specifically, to analyze the drivers of consensus, we estimate the following linear probability model for all participants in the two given treatments:

$$Consensus_2 = \alpha_2 + \beta_2 NotConsensus_1 + \delta_2 X + \gamma_2 X \times NotConsensus_1 + \varepsilon_2 \tag{D.6}$$

where $Consensus_2$ is a dummy variable equal to 1 if a participant selects the same narrative as their partner in the second stage, and 0 otherwise; $NotConsensus_1$ is a dummy equal to 1 if the participant is in a conflicting-narrative pair and 0 if in a same-narrative pair. The coefficient γ_2 represents the *interaction effect*, as reported in Column (2) of Tables D.1, D.2, and D.3. A significant γ_2 indicates that X moderates the relationship between $NotConsensus_1$ and $Consensus_2$.

Similarly, to analyze the drivers of narrative change, we estimate the following linear probability model in the second step:

$$ChangeNarrative_2 = \alpha_3 + \beta_3 NotConsensus_1 + \delta_3 X + \gamma_3 X \times NotConsensus_1 + \varepsilon_3 \quad \text{(D.7)}$$

where $ChangeNarrative_2$ is a dummy variable equal to 1 if a participant changes her narrative in the second stage and 0 if not. The interaction effect γ_3 is reported in Column (3) of Tables D.1, D.2, and D.3. A significant γ_3 means X moderates the relationship between $NotConsensus_1$ and $ChangeNarrative_2$.

Finally, to analyze the drivers of choosing the best-fit narrative, we run the following regression in the second step:

$$BestFit_2 = \alpha_4 + \beta_4 NotBestFit_1 + \delta_4 X + \gamma_4 X \times NotBestFit_1 + \varepsilon_4 \quad \text{(D.8)}$$

where $BestFit_2$ is a dummy variable equal to 1 if a participant chooses the best-fit narrative (based on personal data) in the second stage, and 0 otherwise; $NotBestFit_1$ is a dummy equal to 1 if she did not choose the best-fit narrative based on personal data in the first stage, and 0 otherwise. The interaction effect γ_4 is reported in Column (4) of Tables D.1, D.2, and D.3. A significant γ_4 indicates that X moderates the relationship between $NotBestFit_1$ and $BestFit_2$.

Table D.1: Treatment and interaction effects (*Communication* vs. *Reasoning*)

Explanatory variables	(1)	Interaction effects		
	Treatment effects	(2) Reaching consensus	(3) Changing narrative	(4) Changing to the best-fit narrative
same_ goal	0.117	0.064	-0.014	-0.074
	(0.089)	(0.152)	(0.09)	(0.111)
uncertain	0.147***	0.059	-0.011	0.327*
	(0.044)	(0.205)	(0.171)	(0.180)
agree_ owndata	0.017	0.238***	-0.335***	-0.410***
	(0.012)	(0.079)	(0.046)	(0.057)
diff_ data	0.114	-0.167	-0.163	0.147
	(0.071)	(0.175)	(0.115)	(0.145)
describe_ data	0.069	-0.138	0.102	0.198*
	(0.083)	(0.155)	(0.110)	(0.108)
describe_ method	0.050	-0.154	-0.259	-0.042
	(0.068)	(0.155)	(0.181)	(0.144)
describe_ method1	-0.024	-0.113	-0.278**	-0.192*
	(0.068)	(0.127)	(0.109)	(0.115)
describe_ method2	0.069	-0.155	-0.053	0.280
	(0.044)	(0.131)	(0.147)	(0.175)
describe_ method3	0.119*	0.0165	0.214*	0.134
	(0.066)	(0.0963)	(0.119)	(0.137)
describe_ method4	0.036	-0.144	-0.123	0.077
	(0.047)	(0.176)	(0.216)	(0.213)
ask_ data	0.074	-0.104	-0.016	0.072
	(0.054)	(0.167)	(0.142)	(0.163)
ask_ method	0.016	0.624***	-0.937***	0.610
	(0.02)	(0.221)	(0.224)	(0.442)
real_ unexplained	-0.001	0	0	-0.414***
	(0.017)	(.)	(.)	(0.057)
real_ explained	-0.009	0	0	0
	(0.009)	(.)	(.)	(.)
persuading	0.007	-0.274	-0.349***	-0.415***
	(0.019)	(0.253)	(0.046)	(0.058)
being_ persuaded	-0.029	0.597**	0.196	0.983***
	(0.025)	(0.281)	(0.324)	(0.324)

Notes: Standard errors are shown in parentheses. $n = 232$. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. In column (1), *Communication* serves the benchmark.

Table D.2: Treatment and interaction effects (*Share* vs. *Reasoning*)

Explanatory variables	(1)	Interaction effects		
	Treatment effects	(2) Reaching consensus	(3) Changing narrative	(4) Changing to the best-fit narrative
same_ goal	0.156*	0.075	-0.054	0.049
	(0.088)	(0.149)	(0.0894)	(0.123)
uncertain	0.110**	0.272	-0.182	0.187
	(0.050)	(0.208)	(0.155)	(0.173)
agree_ owndata	-0.002	-0.239	0.129	-0.125
	(0.017)	(0.358)	(0.360)	(0.276)
diff_ data	0.056	-0.098	-0.237*	0.132
	(0.071)	(0.158)	(0.120)	(0.147)
describe_ data	0.449***	-0.338***	-0.054	0.108
	(0.074)	(0.127)	(0.097)	(0.122)
describe_ method	0.209***	-0.307***	-0.305**	0.032
	(0.073)	(0.094)	(0.117)	(0.139)
describe_ method1	0.015	-0.101	-0.205*	-0.098
	(0.067)	(0.121)	(0.104)	(0.126)
describe_ method2	0.085*	-0.170	-0.121	0.292
	(0.045)	(0.140)	(0.147)	(0.198)
describe_ method3	0.167**	-0.086	0.025	0.073
	(0.066)	(0.090)	(0.115)	(0.143)
describe_ method4	0.061	-0.108	-0.415*	-0.103
	(0.043)	(0.175)	(0.212)	(0.215)
ask_ data	0.189***	-0.248*	0.053	0.009
	(0.046)	(0.150)	(0.142)	(0.187)
ask_ method	0.033**	0	0	0.889***
	(0.016)	(.)	(.)	(0.284)
real_ unexplained	0.017	0	0	-0.463***
	(0.012)	(.)	(.)	(0.062)
real_ explained	-0.009	0	0	0
	(0.009)	(.)	(.)	(.)
persuading	0.025*	-0.238	-0.381***	-0.464***
	(0.014)	(0.358)	(0.045)	(0.062)
being_ persuaded	0.016	0.609**	0.293	0.224
	(0.017)	(0.279)	(0.279)	(0.283)

Notes: Standard errors are shown in parentheses. $n = 230$. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. In column (1), *Share* serves the benchmark.

Table D.3: Treatment and interaction effects (*Communication* vs. *Share*)

Explanatory variables	(1) Treatment effects	Interaction effects (2) Reaching consensus	(3) Changing narrative	(4) Changing to the best-fit narrative
same_ goal	0.039	0.112	-0.007	-0.092
	(0.089)	(0.125)	(0.079)	(0.113)
uncertain	-0.037	0.610*	-0.050	0.533*
	(0.037)	(0.317)	(0.223)	(0.269)
agree_ owndata	-0.018	-0.933***	0.596***	0
	(0.013)	(0.065)	(0.034)	(.)
diff_ data	-0.058	0.010	-0.118	0.319**
	(0.066)	(0.134)	(0.124)	(0.150)
describe_ data	0.380***	-0.089	-0.121	0.065
	(0.080)	(0.132)	(0.089)	(0.113)
describe_ method	0.159**	-0.137	-0.108	0.137
	(0.077)	(0.093)	(0.117)	(0.122)
describe_ method1	0.039	-0.015	-0.205*	-0.059
	(0.063)	(0.073)	(0.107)	(0.124)
describe_ method2	0.016	-0.135	-0.125	-0.134
	(0.039)	(0.131)	(0.164)	(0.214)
describe_ method3	0.048	0.016	0.095	0.234*
	(0.063)	(0.082)	(0.115)	(0.139)
describe_ method4	0.026	-0.188	0.097	0.017
	(0.037)	(0.155)	(0.191)	(0.225)
ask_ data	0.116***	0.256	0.119	0.480
	(0.036)	(0.228)	(0.195)	(0.290)
ask_ method	0.018	1.085***	-1.423***	0.546***
	(0.012)	(0.066)	(0.035)	(0.055)
real_ unexplained	0.018	0	0	-0.461***
	(0.012)	(.)	(.)	(0.055)
real_ explained	-0.000	0.078	-0.415***	-0.461***
	(0.013)	(0.067)	(0.036)	(0.055)
persuading	0.018	0	0	-0.461***
	(0.012)	(.)	(.)	(0.055)
being_ persuaded	0.045*	0	0	1.337***
	(0.023)	(.)	(.)	(0.236)

Notes: Standard errors are shown in parentheses. $n = 230$. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. In column (1), *Communication* serves the benchmark.