Silence is Golden: Communication Costs and Team Problem Solving

Gary Charness, UCSB
David J. Cooper, FSU and UEA
Zachary Grossman, UCSB

June 20, 2015

Abstract: Numerous studies have compared the performance of individuals and teams at solving intellective problems. The ubiquitous finding in the economics literature is that teams out-perform individuals. This result is intuitively appealing, as teams can benefit from sharing insights. We analyze experiments comparing the performance of teams and individuals at solving a series of challenging logic puzzles. Contrary to the existing literature, individuals meet or exceed the performance of teams on all measures. If we impose a small cost of communication on teams, the performance of teams improves to closely resemble the performance of individuals. Underlying these results is a definite negative relationship between frequency of communication and team performance. We also document a strong gender effect. Teams with more women perform considerably better even though men slightly outperform women when solving the puzzles individually.

"Silence is golden …"
--The Four Seasons

1. Introduction

Suppose you are a manager facing a difficult problem. You can try to find a solution yourself or you can put together a team to help you. It would seem obvious that you would do better with the help of a team. While there are costs associated with having a team (salaries and opportunity costs), you gain the benefits of the insights of others. Interactions among teammates might even lead to further insights that would not occur to individuals, since diverse ideas can be complementary and build upon each other. Indeed, there is a great deal of research that supports the notion that teams are better than individuals at solving cognitive problems. For example, Charness and Sutter (2012) document that teams in economics experiments are considerably better at accomplishing such tasks in a wide array of games.[1]

However, anyone who has been on a committee may be less convinced about the effectiveness of teams or the value of communication from co-workers. Many or most readers have been in meetings that were an enormous waste of time or had a co-worker who was greatly enamored with the sound of his or her own voice. Not all shared insights are good insights and it takes time to separate good ideas from bad ones. Is it truly obvious that having a team is worth the costs? Is increased communication necessarily a good thing?[2]

This paper presents a series of experiments that provide negative answers to both of these questions. Subjects are confronted with a series of challenging logic problems. They attempt to

[1] Examples from the psychology literature include Shaw (1932), Laughlin, Bonner, and Miner (2002), Tindale, Kameda, and Hinsz (2003), and Kugler, Kausel, and Kocher (2012).

[2] A related literature notes that having too large a group is counter-productive. While more people in a group increases the likelihood that someone will propose the correct decision, this also means that there are more opinions and ideas that must be communicated and discussed. Hackman and Vidmar (1970) asked participants who had performed group tasks of various sorts to indicate the optimal group size (this was between four and five. See also Bray, Kerr, and Atkin (1978) and Blenko, Mankins, and Rogers (2009). Also related is research showing that teams often fail to beat the demanding "truth wins" benchmark in logic problems (Lorge and Solomon, 1955) and strategic environments (Cooper and Kagel, 2005; Casari, Zhang, and Jackson, 2014; Cooper and Sutter, 2014).

solve these puzzles either as individuals acting alone or in groups of four people working together. In the latter case, one subject is the “leader,” filling out the puzzle, while the other three teammates are “followers” who assist the leader. Followers have the same information as leaders, seeing all entries into the puzzles in real time, and receive the same payoff as the leader. Unlimited free-form chat allows teammates to easily share insights on how to solve the puzzles.

Relative to the performance of individuals acting alone, teams do *not* perform better. If anything, teams perform *worse* than individuals since individuals are on average significantly more likely to solve the puzzle rapidly.[3] We find strong evidence of congestion effects, as sending more messages reduces performance on all measures. This suggests that the relatively poor performance of teams can be improved by reducing the amount of communication.

A natural approach to limiting the number of messages is to add a cost for sending messages. We test whether imposing a tiny cost (a penny each) for sending messages improves team performance. Although messages are quite inexpensive, there is a dramatic decrease in the number of messages sent. Limiting communication improves group performance on all measures, particularly in terms of the likelihood of a quick solution. To the best of our knowledge, our study is the first to find that imposing friction on communication leads to more effective performance in teams.

The positive effect of adding a message cost is smaller than expected given the large reduction in how many messages are sent. This is due to an unanticipated side effect of message costs: subjects respond to the increased message cost by cramming many more suggestions into each message. Consistent with messages being denser, the marginal negative effect of a message is significantly larger with message costs. Getting rid of congestion effects is surprisingly difficult because subjects are good at finding ways around an increased cost of messages.

[3] The effect of having a team is heterogeneous as low-ability individuals do better as leaders than as individuals.

In addition to the main results about the effects of teams and communication costs, we find a strong gender result. While females as individuals do a bit worse than men at solving puzzles, groups with a majority of women significantly outperform those with a (weak) majority of men. Perhaps women are better at the communication process or have superior team skills.

Our results have applications to contemporary settings, in which a great deal of time is wasted on excessive business-related e-mail messages, meetings, and committees. We suspect that most of our readers would gladly endorse a call for fewer emails, meetings, and committees. More generally, it is easy for one to feel bombarded by information that is conveyed with no cost. Our results sound a cautionary note regarding the notion that cheap-talk messages are always highly effective or at least harmless. They also indicate that it may be surprisingly difficult to shut down wasteful communication.

The remainder of the paper is as follows. Section 2 is a literature review and we present our experimental design and implementation in Section 3. Results and analysis follow in Section 4, and we offer some discussion in Section 5. Section 6 concludes.

2. Literature Review

There is a large literature demonstrating that teams outperform individuals in cognitive tasks. In psychology, the earliest evidence of which we are aware is Shaw (1932), who found that groups were seven times as likely to solve puzzles correctly, with the group advantage stemming largely from the checking of errors and the rejection of incorrect solutions. Lorge and Solomon (1955), another early article in psychology, originated the "truth-wins norm." The idea is that in "eureka-type" problems, where there is a solution that is transparent once seen or explained, a group should do as well as its most able individual since this person can solve the problem and explain it to others.[4] The psychology literature finds that groups rarely meet and almost never exceed the truth-wins norm when solving logic problems. This failure is attributed to "process loss," a broad term that incorporates both free-riding and congestion effects.

Charness and Sutter (2012) present a detailed summary of the economics literature on group decision-making. The main finding is that groups almost invariably make better self-interested decisions than individuals do. A number of researchers have studied performance relative to the truth-wins norm for teams playing games. Cooper and Kagel (2005, 2009, 2015) find that teams in a difficult signaling game consistently play more strategically than individuals and beat the truth-wins norm in more difficult games. While Cooper and Sutter (2014) and Casari, Zhang, and Jackson (2014) find that teams outperform individuals, their results show that groups fail to beat the truth-wins norm in takeover games. The difference appears to stem from whether the subjects perceive there is a central insight that can be passed on to others.[5]

There is at least some work on the effect of gender composition on team performance. Wooley, Chabris, Pentland, Hashmi, and Malone (2010) find that the group's collective intelligence is correlated with the proportion of females in the group. Fenwick and Neal (2001) showed that groups with the same number of men and women out-performed homogenous groups, and Apesteguia, Azmat, and Iriberri (2012) and Hoogendoorn, Oosterbeek, van Praag (2013*)* find that mixed-gender groups make better decisions than do same-gender teams. However, a study of a Fortune 500 firm in the information processing reported in Kochan *et al.* (2003) found no effects for team-level gender diversity on team performance. Our study differs from these studies in that they look at business simulations in which teams must perform many

[4] For example, if each person is 50 percent likely to see a solution and the probability across individuals is uncorrelated, the likelihood of solution is 75 percent with two people, 87.5 percent with three people, etc. Note that the marginal gain from adding an additional person becomes smaller and smaller, while coordination issues grow.

[5] In this vein, Isopi, Nosenzo, and Starmer (2011) find that teams do worse than individuals in a task with no demonstrably correct solution. Casari *et al.* (2014) report a similar result for one of their treatments.

types of tasks, while we focus on one specific laboratory task with controls and incentives. We find that the results improve monotonically with the number of women in the group.

We have known for decades that cheap-talk (i.e. costless) communication can yield more efficient outcomes. There are many prominent examples within economics for social (and individual) dilemmas, trust games, and coordination games. Recent papers have made progress in understanding how and why this works (simple vs. free-form messages, guilt aversion, coordinating on an equilibrium/punishment scheme).

Simple, categorical messages typically successfully achieve payoff-dominant outcomes in coordination games, which feature multiple pure-strategy equilibria. The first such demonstration was Cooper, DeJong, Forsythe, and Ross (1992), finding strong coordination in the second half of their sessions with two-way communication. Other papers, including Charness (2000), Blume and Ortmann (2007), and Brandts and Cooper (2007), find similar effectiveness. However, in games with a unique and socially-inefficient equilibrium, such simple messages have been found to be ineffective, although not harmful; examples include studies by Charness (2000), Charness and Dufwenberg (2010), Ben-Ner, Putterman, and Ren (2011), Andreoni (2014), and Oprea, Charness, and Friedman (2014).[6]

Nevertheless, anonymous *free-form* messages have been used with great success. Charness and Dufwenberg (2006, 2011) observe large improvements in social efficiency (total payoffs) in two-person sequential games when the second mover is permitted to send an endogenous written message to the first mover; the authors attribute this to changes in beliefs. Brandts, Charness, and Ellman (forthcoming) find that free-form messages not only lead to

[6] Brandts, and Cooper (2007) and Brandts, Cooper, and Weber (2014) find better results for teams in coordination games with communication than with heightened incentives. Charness, Karni, and Levin (2007, 2010) find that consultation with others improves choices relative to no consultation.

Pareto-improvements in payoffs for buyers and sellers, but they also even change the predominant form of contract. Cooper and Kühn (2014) find that free-form communication increases total surplus in a repeated Bertrand oligopoly by generating coordination on an equilibrium which includes punishment of deviations from cooperative play.

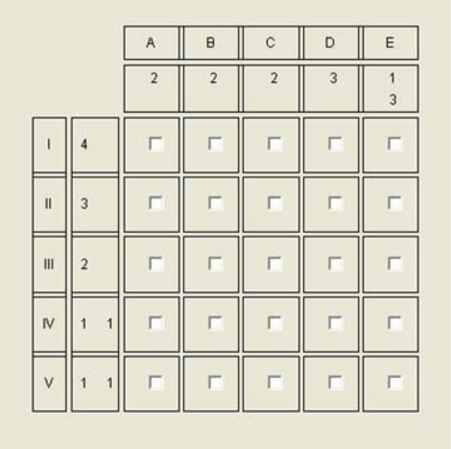
There has been almost no work on costly communication in experiments. Blume, Kriss, and Weber (2014) study stag-hunt games. They find that imposing modest costs for sending messages reduces the use of messages, but efficient coordination occurs with similar frequency as when there are no costs. Their results are consistent with a formalization of forward induction that selects the efficient pure-strategy equilibrium outcome without communication. Wilson (2014) tests a model of group-based deliberation where both sending and receiving messages is costly. He finds excess communication when costs are high, which could have improved welfare, since the information in this setting is a public good. However, subjects actually do worse than predicted in equilibrium, because they use the information sub-optimally. Thus total welfare (net of message costs) may be reduced by the existence of costly communication channels.

We are unaware of any previous experimental study in which costly communication improves outcomes. Our messages have very small costs and so one would not particularly expect behavior to be greatly changed. Nevertheless, we find that not only do message costs lead to a dramatic reduction in the number of messages sent, they lead to better outcomes for the groups on all measures, significantly so for the likelihood of a rapid solution.

3. Experimental Design

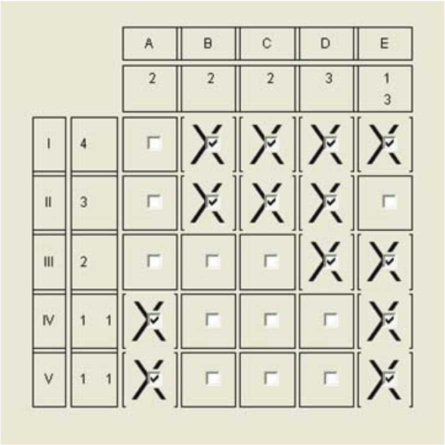
Sessions began with instructions, which can be found in Appendix C. These were read aloud while the participants followed along and were allowed to ask questions. Participants tried to solve grid-based logic puzzles called *nonograms*. This served as a real-effort task that was challenging, yet with rules simple enough to be learned quickly. Because they require many steps of reasoning to solve, they provide many opportunities for groups to communicate and work together. We selected these puzzles, instead of a more familiar alternative such as Sudoku, so that most participants would enter the session with little previous experience.

Figure 1a: Unsolved Nonogram



		A	B	C	D	E
		2	2	2	3	1 3
I	4	□	□	□	□	□
II	3	□	□	□	□	□
III	2	□	□	□	□	□
IV	1 1	□	□	□	□	□
V	1 1	□	□	□	□	□

Figure 1b: Solved Nonogram



		A	B	C	D	E
		2	2	2	3	1 3
I	4	□	☒	☒	☒	☒
II	3	□	☒	☒	☒	□
III	2	□	□	□	☒	☒
IV	1 1	☒	□	□	□	☒
V	1 1	☒	□	□	□	☒

Figure 1a shows a screen-shot of an unsolved 5 x 5 nonogram from the computer interface used in the sessions. Each cell can be marked or left unmarked (the dot in the center of each cell is a radio button that participants clicked to mark and unmark cells). The goal is to correctly determine which cells should be marked. Once subjects thought they had solved the puzzle, they needed to click the button in the lower right corner labeled "Check Answer" to submit their solution. If the puzzle had not been solved correctly, the subject(s) saw a message telling them their solution was incorrect. There was no penalty for submitting an incorrect solution and work on the puzzle could continue.[7] For reference, Roman numerals index rows and letters index columns. Labels with Arabic numerals indicate the length of each run of

[7] There was also a button, labeled "Retry (clear board)" that allowed subjects to return the puzzle to its original unmarked state.

consecutive marked cells, according to the solution, in each row and column. Figure 1b shows the same puzzle, correctly solved, as it appears on the computer interface.

Each session proceeded in two stages, with five rounds of puzzles in each stage. We used the same puzzles in the same order for all sessions.[8] Stage 1 consisted of puzzles that were relatively small (5 x 5) and therefore easier. Participants worked on them individually with a $0.50 incentive for correctly solving the puzzle within a 95-second time limit. The purpose of this stage was to give participants an opportunity to familiarize themselves with the rules and strategy as well as to provide a measure of individual ability. Participants averaged 2.57 correct solutions across the five rounds of Stage 1, with a wide spread in performance (StDev = 1.61).

Stage 2 featured larger (10 x 10), and therefore more difficult, puzzles with a nine-minute time limit. Three main experimental treatments determined whether participants worked on these puzzles alone or in teams, and how expensive it was for teammates to communicate. As secondary treatments, we also varied the incentives for solving the puzzles.

In Stage 2 of the *Individual* treatment, participants worked on the puzzles and were incentivized independently. This treatment served as a control for the other two treatments, in which participants worked on the puzzles and were incentivized as groups. Participants did not interact with each other in the *Individual* treatment, so each individual is an independent observation. This let us use fewer subjects and sessions than in the team treatments.

In the team treatments (*Team-Cost* and *Team-No Cost*), participants were assigned to groups of four consisting of one leader and three followers. Subjects stayed in the same role, leader or follower, for all five rounds of Stage 2. Groups were re-assigned each round using a stranger matching protocol and participants did not know the identities of their teammates.

[8] The only exception to this is the puzzle used in round 10 of the first session. We deemed this puzzle too difficult for our purposes (no group solved it or even came close) and replaced it with an easier puzzle in all subsequent sessions. We omit data using the original round-10 puzzle from our analysis.

Group members could send typed messages to each other through a chat box.[9] The messages were labeled with an ID number making it possible within a round to tell which group member had sent which message. ID numbers (1, 2, 3, or 4) were randomly assigned within each group in each round. The instructions stressed that ID numbers were redrawn in each period, so participants knew they could not use the ID numbers to identify individuals across rounds. Only the leader could directly work on the puzzle, choosing which cells to mark or unmark and submitting solutions. The followers could see the current state of the puzzle – this was reflected on followers' screens in real time as the leader marked or unmarked cells – but were limited to advising the leader via chat messages.

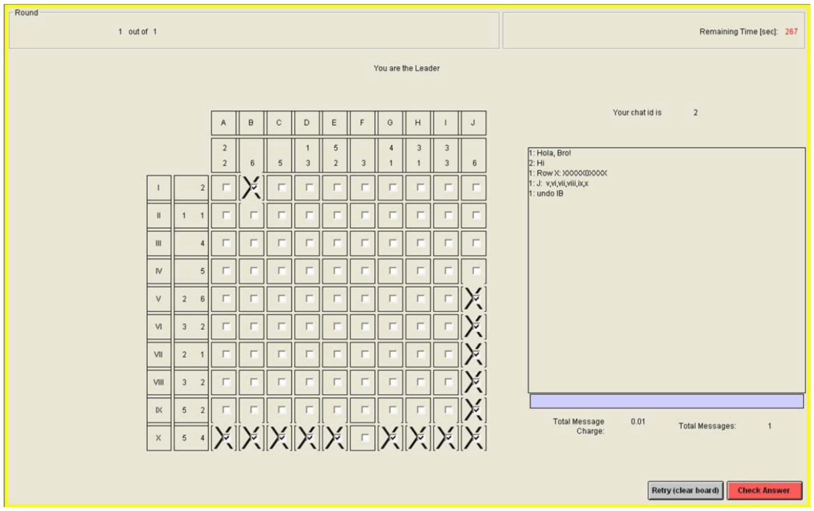
We were concerned that there would be little communication, even without message costs, due to leaders attempting to solve the puzzles on their own. We therefore initially biased the selection of leaders towards lower-ability individuals, as measured by the number of puzzles solved in Stage 1, to encourage communication. In the earlier team sessions, the weakest performers were assigned the leader roles (ties were resolved randomly). In practice, there was no shortage of communication even with high-ability leaders, and so for the remaining team sessions all participants were equally likely to be assigned the leader role. In no case did the instructions give any information about assignment to leader or follower roles.

Participants in the *Team-No Cost* treatment could send messages at no monetary cost, while participants in the *Team-Cost* treatment paid $0.01 for each message sent. We chose this cost to be very small relative to the size of the reward for completing the puzzle, with the idea that it would deter marginal messages but not eliminate communication altogether. The *Team-Cost* treatment was otherwise identical to the *Team-No Cost* treatment. Figure 2 shows a screen-

[9] The subjects were told that they could use the chat box to "advise each other." The only specific restrictions we put on communication were telling them not to identify themselves and to avoid offensive language.

shot for a leader in the *Team-Cost* treatment. The puzzle is on the left and the chat box is on the right with a running summary of message costs below it.

Figure 2: Leader's Screen from Team-Cost Treatment



Round

1 out of 1

Remaining Time [sec]: 267

You are the Leader

		A	B	C	D	E	F	G	H	I	J
		2 2	6	5	1 3	5 2	3	4 1	3 1	3 3	6
I	2	☐	☒	☐	☐	☐	☐	☐	☐	☐	☐
II	1 1	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
III	4	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
IV	5	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
V	2 6	☐	☐	☐	☐	☐	☐	☐	☐	☐	☒
VI	3 2	☐	☐	☐	☐	☐	☐	☐	☐	☐	☒
VII	2 1	☐	☐	☐	☐	☐	☐	☐	☐	☐	☒
VIII	3 2	☐	☐	☐	☐	☐	☐	☐	☐	☐	☒
IX	5 2	☐	☐	☐	☐	☐	☐	☐	☐	☐	☒
X	5 4	☒	☒	☒	☒	☒	☐	☒	☒	☒	☒

Your chat id is 2

1: Hola, Bro!
2: Hi
1: Row X: XXXXX0XXXX
1: J: v,vi,vii,viii,ix,x
1: undo IB

Total Message Charge: 0.01

Total Messages: 1

Beyond the three main treatments, two dimensions of the monetary incentives were varied across sessions. Both were balanced across treatments. In *Low Pay* sessions, groups or individuals who failed to solve the puzzle within the time limit received $0, while groups or individuals who were successful each earned $3 per person.   In *High Pay* sessions, unsuccessful individuals or group members received $1 each, while success earned $5 per person. Thus, both the total and marginal benefits of finishing the puzzle were higher in the *High Pay* sessions.[10]

The second incentive variation took the form of a time bonus. In the *Time Bonus* sessions, members of groups that completed the puzzle earned an additional $0.01 for every second that remained before the time limit when they finished. In the *No Bonus* sessions there was no incentive pay for solving the puzzles rapidly.

All individuals in a group received the same pay (and the same time bonus) for solving the puzzle; pay could vary within a group due to charges for the number of messages sent.

[10] Having *Low Pay* and *High Pay* reflects the history of the experiments and was not intended to study any particular hypothesis. The response to this variation is not central to our hypotheses and we treat it as a nuisance variable.

Sessions of the *Individual* treatment were also balanced between *High* and *Low Pay* and whether there was a time bonus. The time bonus was calculated in the same manner as for teams.

Table 1: Summary of Sessions

	No Time Bonus		Time Bonus	
	Low Pay	High Pay	Low Pay	High Pay
Individual	18 subjects 1 session	16 subjects 1 session	15 subjects 1 session	19 subjects 1 session
Team – Cost	48 subjects 2 sessions	40 subjects 2 sessions	40 subjects 2 sessions	40 subjects 2 sessions
Team – No Cost	56 subjects 3 sessions	60 subjects 3 sessions	36 subjects 2 sessions	44 subjects 2 sessions

We conducted 22 sessions in the xs/fs laboratory at Florida State University, each lasting 75 to 90 minutes. Table 1 summarizes the number of sessions and subjects used.[11] The data includes observations from 68 participants in the *Individual* treatment, 168 participants in the *Team-Cost* treatment, and 196 participants in the *Team-No Cost* treatment. Sessions had between 16 and 24 participants who were recruited with the software ORSEE (Greiner 2004) and participated via a computer interface programmed in Z-Tree (Fischbacher 2007). Participants were separated by privacy dividers and were not allowed to talk to each other except through the chat program. Earnings were the sum of a $10 show-up fee and accrued earnings across the ten puzzle rounds. Total earnings, including the show-up fee, averaged roughly $27.

4. Hypotheses

These experiments study whether having a team always improves performance and whether increased communication necessarily helps team performance. This section proposes

[11] Data from Round 10 in the first session was dropped due to puzzle difficulty and data from Round 6 in one session was dropped due to a software error. We ran replacement sessions in both cases, yielding an extra session in two cells. One person walked out of a session of the *Individual* treatment, leaving 15 subjects in that session.

several hypotheses about the experimental results, derived partially by developing a simple model of behavior.

All of the hypotheses relate to puzzle-solving performance. We evaluate each hypothesis using two performance measures: the percentage of puzzles solved correctly and percentage of fast solutions, defined as solving the puzzle in less than half of the available time. While the frequency of correct solutions is a critical issue and is the primary determinant of earnings in our experiment, speed is equally (or more) important in determining payoffs in many organizational settings, particularly in terms of cost-effectiveness. This is imposed directly in our time bonus treatment, but applies in any setting where profit is a function of how many tasks an individual or group completes in a given time period.[12]

Also note that our 9-minute deadline for the 10x10 puzzles was arbitrary and imposed due to logistical considerations. Perhaps many individuals or groups who were unable to solve a puzzle in 9 minutes would have solved it with more time. Likewise, there were presumably individuals and groups who solved the puzzle who would not have done so given less time. As a performance measure, speed is less sensitive to the specific time limit we imposed.

In Section 2 we discussed the literature on team problem-solving. Our first two hypotheses are based on common findings in this literature: freely interacting teams generally perform better than individuals but fail to meet the demanding truth-wins norm.

H1: Teams without message costs will perform better than individuals.

To understand what the truth-wins norm implies for our experiment, consider the following way a group might operate (this is *not* what happens in our experiment). Each group member is given a copy of the puzzle and works on it independently. When one group member

[12] As an illustration of why speed is an important performance measure, think about the problem facing a professor trying to get tenure. You have a fixed period of time and are rewarded not just for publishing a single paper, but for the number of papers published. Speed matters.

solves the puzzle, the entire group is considered to have solved the puzzle. The group's speed of solving the puzzle is determined by the speed of the most able group member. If having a freely-interacting team generates positive synergies, a team should do better than its best member.

H2: Teams without message costs will perform as well as or better than the best of four randomly-selected subjects from the Individual treatment.

The next two hypotheses flow from a simple model of a follower's decision about whether or not to send a message. Here we describe the key insights in intuitive terms, while the full model is presented in Appendix A. The basic premise is that a follower with an insight will send a message if the perceived benefit outweighs the cost. The benefit of sending a message depends upon the reward for solving the puzzle and the extent to which sending the message increases the probability of doing so. The cost of sending a message includes both any potential monetary costs, such as those we impose in the *Team-Cost* treatment, and possibly non-pecuniary costs. The benefits accrue to all group members but the follower considers only her private benefit, so for many parameter values this model mirrors a standard public-good (or joint-production) problem. In such cases, sending a message generates a positive externality and a self-regarding follower sends too few messages from a social point of view.

The critical insight is that this simple model does not always yield a standard public-goods game. It is easy to devise cases in which sending a message generates a negative externality. An obvious one is when the follower's perception of the value of her message is mistaken. She may think that she is sending a helpful message when it is actually harmful (e.g., incorrect advice or irrelevant chatter that distracts other group members). Or she may simply overestimate how helpful her message is. While a follower might under-estimate the value of her message, the over-estimation case is more interesting because it admits the possibility that messages carry a negative externality and that followers will send more than is socially-optimal.

Similarly, non-pecuniary costs can take a range of values. Ideas do not arrive fully formed and it may take some effort to bring them to fruition.[13] Sending a message may also generate utility or disutility independent from the effort spent on producing it. A shy individual may experience disutility from sending a message, while others may enjoy sending a humorous message. Even when a follower understands that her messages do not help, if this positive benefit (negative cost) of sending a message is sufficiently strong, it may be privately optimal for her to do so. This too will lead her to send too many messages from a social point of view.

The model points to the central empirical issue in this paper. The value of messages and, by extension, the effect of increasing message costs hinges on whether followers' messages are actually useful for solving the puzzle. If there is a positive externality associated with sending messages, adding a monetary message cost, as occurs in the *Team–Cost* treatment, will exacerbate the problem and reduce performance. If there is a negative externality, it is helpful to limit the number of messages sent and performance will improve. Evaluating the following pair of hypotheses will provide an empirical answer to whether or not this is the case.

H3: Ceteris paribus, groups that send more relevant messages will perform better.

H4: Group performance will be lower in the Team – Cost treatment than in the Team – No Cost treatment.

H4 depends on H3 being true. If messages tend to be beneficial, cutting the number of messages is harmful. It is not a given that messages are typically beneficial. If messages harm performance, cutting the number of messages would be helpful.

We expected the time bonus to improve performance, both by increasing effort in general and by reducing the number of irrelevant comments. Knowing that time was, literally, money,

[13] There is also an effort cost in typing or reading a message, although it is most likely low for people raised on chat.

subjects would not want to waste time by either sending or forcing others to read messages that were obviously off-task. This implies better performance, especially regarding solution time.

5. Results

A. An Overview of the Results: The purpose of Section 5A is to give a sense of the data, with serious statistical analysis postponed until Section 5B.

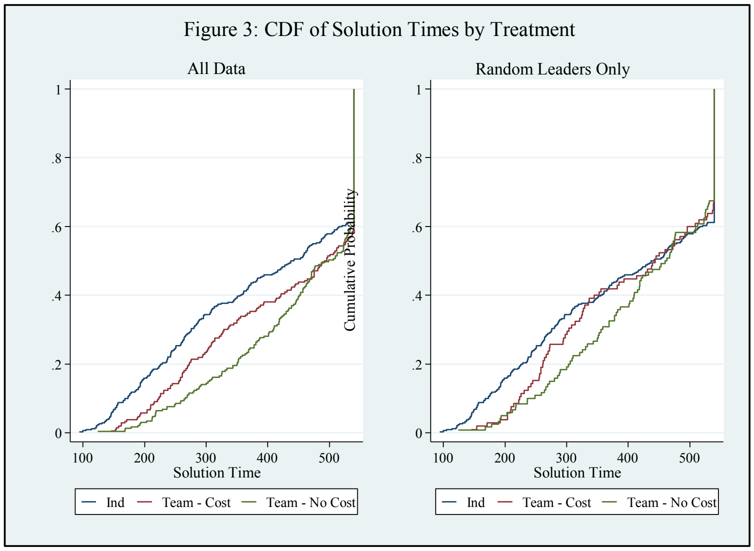


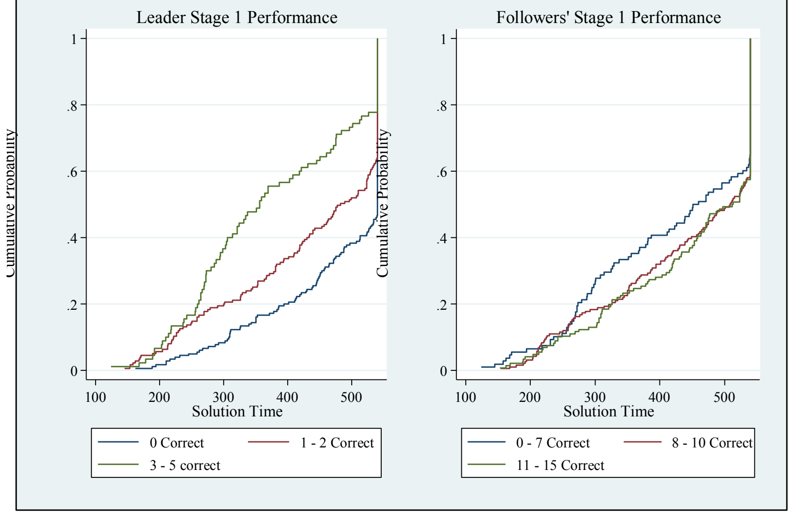
Figure 3: CDF of Solution Times by Treatment

Figure 3 displays the CDF of solution time broken down by the main treatments. Groups that did not solve the puzzle are assigned a solution time of 540 seconds, so the colored vertical line at the upper right of each panel shows groups that failed, not a burst of solutions at the last second. The left panel is based on all of the data, while the panel on the right restricts the dataset for teams to sessions where the leaders were selected randomly. In the left panel, the line for the *Individual* treatment is always above the line for the *Team-Cost* treatment, which likewise is almost always above the line for the *Team-No Cost* treatment. The differences between the treatments narrow at the end of the 540 seconds. Individuals are more likely than teams (without message costs) to solve the problem quickly – 29 percent of individuals solve the puzzle in less than half the available time versus 11 percent of teams– but solution rates after the full nine minutes are almost identical across the treatments, ranging from 61 to 62 percent.[14] Neither H1 nor H4 receives much initial support from the data. Given that the truth-wins norm is more demanding than merely asking teams to out-perform individuals, it follows that H2 also fares poorly.

Leaders are chosen from the lowest performers in Stage 1 for half of the team sessions, potentially biasing the comparison between teams and individuals. The right panel of Figure 3 confirms the existence of this bias and justifies the need for the regressions (reported in Section 5B) that control for leader and follower ability. The graph is noisier due to having fewer observations, but the difference between teams and individuals narrows and disappears around the 400-second mark. The order over treatments remains the same.

Comparing the two panels of Figure 3 suggests that leaders matter more than followers. Figure 4 directly examines how team performance varies with the ability of the leaders and the followers. Data is taken from the team treatments. Subjects faced a series of five relatively easy nonograms in Stage 1. We use the number solved as a measure of individual ability.

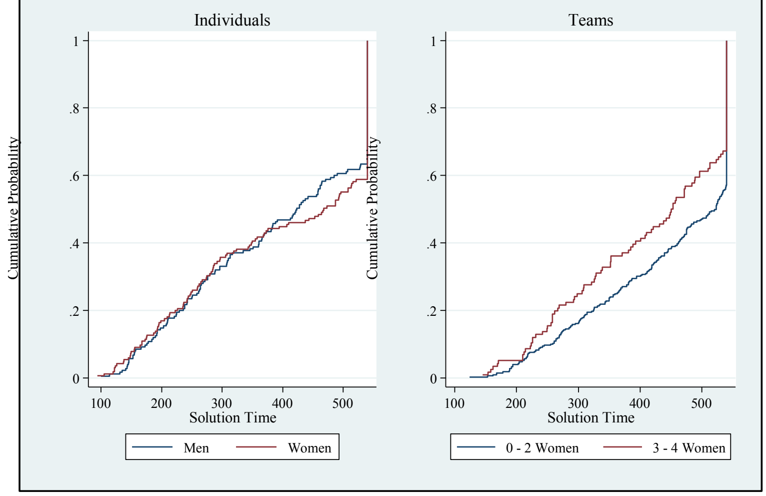
Figure 4: Stage 1 Performance and CDF of Solution Times



[14] Note the lack of a deadline effect, as the rate of individuals/groups does not increase near the end of the available time. There is no evidence that people increase effort or otherwise speed up as the time limit approaches.

In the left panel, the data are broken down by how many Stage 1 problems were solved by the group leader: 0 correct (180 obs.), 1-2 correct (175 obs.), or 3-5 correct (90 obs.). As expected, there is a strong positive relationship between the number of Stage-1 problems solved by the leader and group performance in Stage 2. The right panel breaks down the data by how many Stage-1 problems were solved by the three followers: 0–7 correct (108 obs.), 8–10 correct (191 obs.), or 11–15 correct (146 obs.). The relationship between followers' Stage-1 performance and group performance in Stage 2 is weak and has an unexpected sign. The data are inconsistent with H2, and also inconsistent with the model of team processes underlying H2. The leader matters enormously for team performance, but the followers are largely irrelevant.

Figure 5: Gender and CDF of Solution Times



Our experiments were not designed with gender effects in mind. Nevertheless, while we had no real *ex-ante* hypothesis, we gathered data about gender as a matter of course. Figure 5 displays the effect of gender on performance. The left panel shows data from the *Individual* treatment; men slightly out-perform women. The right panel is based on data from the team treatments, subdivided by teams with a (weak) majority of men and teams with a majority of women. Although women have no more inherent talent for solving these puzzles, teams with a majority of women strongly out-perform teams with a (weak) majority of men.

B. Regression Analysis: The regressions presented in this section provide formal statistical support for our conclusions. This is particularly important because the selection process was intentionally biased to assign more poor performers from Stage 1 to the leader role. Since the leader's ability (measured by Stage-1 performance) is more important than the followers' ability, over-sampling low-ability leaders biases performance downward. As suggested earlier, the relatively poor performance of teams relative to individuals could potentially reflect this bias, rather than or in addition to generally poor performance by teams.

The regressions reported in Table 2 correct for the assignment process. We study correct and fast solutions (less than 270 seconds, half the available time). The regressions are probits given that the dependent variables are binary outcomes. Standard errors (in parentheses) are corrected for clustering. A cluster is defined as observations from the same individual for the *Individual* treatment and from the same session for the team treatments (observations from the same session are not independent due to the random re-matching).

In all of the regressions, the base (i.e. the omitted category) is the *Team-No Cost* treatment with *Low Pay* and no time bonus. Dummies for the other two main treatments (*Individual* and *Team-Cost*) capture differences from the *Team-No Cost* treatment.[15] All regressions include dummies for the two payment variations (*Time Bonus* and *High Pay*) as well as controls for Stage 1 performance. For the *Individual* treatment, there are no followers. We therefore de-mean the Stage 1 performance measure for followers to avoid biasing the estimated

[15] None of our hypotheses compare the *Individual* and *Team – Cost* treatments, but for completeness we report estimates for the difference between these treatments at the bottom of the table. These are not a parameter of the model, but rather are derived from the parameters for *Individual* and *Team – Cost*.

effect of the *Individual* treatment.[16] Models 2a and 2b add controls for gender in the *Individual* treatment and the number of women in the group for the team treatments. We de-mean the latter variable to avoid biasing the estimated treatment effect for *Individual*.

Table 2: Regression Analysis of Treatment Effects

Dependent Variable	Correct Solution		Fast Solution (Solution Time < 270 seconds)	
Model #	Model 1a	Model 2a	Model 1b	Model 2b
Individual	-0.266 (0.168)	-0.290* (0.163)	0.462** (0.192)	0.424** (0.191)
Team – Cost	0.073 (0.186)	0.047 (0.178)	0.448** (0.200)	0.382** (0.191)
Time Bonus	0.152 (0.136)	0.138 (0.135)	0.063 (0.134)	0.051 (0.135)
High Pay	0.228 (0.139)	0.221 (0.137)	-0.000 (0.138)	0.004 (0.137)
Leader Stage 1 Correct	0.337*** (0.054)	0.349*** (0.052)	0.276*** (0.041)	0.293*** (0.040)
Followers Stage 1 Correct (DM)	0.043 (0.037)	0.044 (0.036)	0.001 (0.029)	0.002 (0.031)
Women Individual	-	0.089 (0.214)	-	0.208 (0.183)
# Women Teams (DM)	-	0.179*** (0.061)	-	0.191* (0.103)
Individual (vs. Team – Cost)	-0.339 (0.212)	-0.337* (0.202)	0.014 (0.150)	0.042 (0.152)
# Observations	785	785	785	785

* p<0.1; ** p<0.05; *** p<0.01, two-tailed tests

We begin by evaluating whether having a team is helpful, as per H1 and H2. Individuals are less likely than teams to solve the puzzle without message costs, but the effect only becomes weakly significant with the addition of gender controls. Individuals are *more* likely to solve the puzzles quickly than teams without message costs. The estimated effect is large (for Model 1b,

[16] Specifically, set "Group Stage 1 Correct" equal to zero for the *Individual* treatment and, for data from the two team treatments, sum the number of problems solved in Stage 1 by the three followers and demean by subtracting three times the average number of problems solved in Stage 1 for all individuals.

the implied likelihood of a fast solution is 12 percentage points higher for individuals) and significant at the 5% level with or without gender controls.

Conclusion 1: Controlling for the ability of subjects, the data provide no support for H1 or H2, as teams without message costs do not perform better than individuals.

From a manager's point of view, hiring a team of helpers only justifies the cost if it significantly improves performance. The preceding result suggests that any performance gains from having a team are minimal, but this masks underlying heterogeneity. If we look at only low-ability individuals (less than two puzzles solved in Stage 1) and teams with low-ability leaders, teams without message costs have higher solution rates than individuals (58 percent versus 30 percent) and about the same chance of a fast solution (eight percent versus seven percent). Re-running Model 1a with this subsample, the difference in solution rates is significant at $p = 0.001$.[17] Having a team of followers makes sense for low-ability individuals.

Based on the simple model we hypothesized that message costs would harm team performance, but the regression analysis does not support this hypothesis. The *Team-Cost* coefficient is positive across all models. Teams with message costs are slightly more likely to solve the puzzles, as shown in Models 1a and 2a, but the effect is small and not significant. Adding message costs has a much larger effect on the likelihood of a fast solution. The estimated effect is large (for Model 1b, the implied likelihood of a fast solution is 12 percentage points higher for teams with a message cost than without) and significant at $p = 0.013$ or better with or without gender controls.

Conclusion 2: Controlling for the ability of subjects in the various roles, we find no evidence that adding message costs harms the performance of teams. The data do not support H4.

As a measure of solution speed, Model 1b uses whether or not the problem was solved in less than 270 seconds (half the available time). This definition of a fast solution is arbitrary, but

[17] The parameter estimate for the Individual treatment dummy is -0.279, with a standard error of 0.082.

we can check whether our conclusions are robust by systematically varying the cutoff used to define a fast solution. Doing so and re-running Model 1b for alternative definitions of a fast solution, we find that teams with message costs are significantly more likely to solve the puzzles for cutoffs at or below 400 seconds (see Appendix B for details). This effect diminishes (and is not statistically significant) for higher cutoffs, so teams without message costs eventually catch up to those with message costs, as is shown in Figure 3.

As another way of looking at how the treatments affect solution speed, we have run double-hurdle models equivalent to Models 1 and 2 above. The results of these regressions measure whether the treatments affect the solution speed conditional on solving the puzzle. Both the *Individual* and *Team-Cost* treatments lead to significantly faster conditional solution speeds relative to *Team-No Cost*. For Model 1, the estimated improvements are 84.4 and 41.8 seconds, respectively.[18] We will rely primarily on the probits like those shown in Table 2, since these are simple and make it possible to implement IV regressions later in the paper, but our conclusions are robust to how we measure the effects of the treatments on solution speed.

Adding a bonus for solving the problem early has a positive but insignificant effect on performance. The strongest effect of adding message costs is on speed of solutions. This effect does not depend on whether or not there is a time bonus. If we re-run Model 1b replacing the dummy for *Team-Cost* with two separate dummies for *Team-Cost* with and without the time bonus, both parameter estimates are significant. The estimate with the time bonus is a bit larger, but the difference is not significant.[19] The lack of an effect from the time bonus implies that the effect of messages costs on speed does not depend on subjects having an incentive to solve the puzzle quickly. As will be shown below, subjects were largely on task and sent relevant

[18] The standard errors are 28.1 and 27.8. The estimates are almost the same in Model 2 at 83.7 and 40.3 seconds.

[19] The parameter estimates for Team-Cost with and without the time bonus are 0.406 and 0.490 respectively, with standard errors of 0.213 and 0.274.

messages. Making messages costless affects the speed of solutions by changing the process by which puzzles are solved, not by inducing subjects to intentionally delay solving the puzzles.

Conclusion 3: Controlling for the ability of subjects in the various roles, we find no evidence that adding a time bonus improves performance and we find no significant evidence of a difference across high pay and low pay.

The regressions confirm that the performance of teams was very sensitive to the performance of leaders but not followers. In all four models the parameter for the leader's Stage 1 performance is large and significant at $p = 0.01$, while the parameter for the followers' Stage 1 performance is small and never significant.[20]

Female gender has a weak positive effect on performance in the Individual treatment,[21] but a strong and significant one in the team treatments. The effect is stronger for correct solutions than fast solutions, which is consistent with Figure 5 where the gender effect widens over time. Looking at interactions with role (leader or follower), it does not matter if the women are in one role or the other.[22] The effect with teams may not appear that much bigger than those reported for the gender dummy in the *Individual* treatment, but the range is 0-4 rather than 0-1. Controlling for Stage-1 performance, a woman in the *Individual* treatment is estimated to be 3.3 percentage points more likely than a man to solve a puzzle. A team with four women is 26.8 percentage points more likely than a team with four men to solve a puzzle!

Conclusion 4: The performance of teams is increasing in the ability of the leader and in the number of women, but not in the ability of the followers.

[20] As an alternative, we have run regressions where the responsiveness to Stage 1 performance is fit separately for individuals and leaders. This has no qualitative effect on our conclusions. The leader's ability has a strong and significant effect while followers' abilities are almost irrelevant.

[21] The positive estimate is due to controls for Stage 1 performance. While performance by men and women is virtually identical in Stage 1 (53 percent versus 49 percent correct solutions), the women drawn for the *Individual* treatment did worse than the men (54 percent vs. 39 percent). If we don't control for Stage 1 performance, the estimated gender effect for the Individual treatment is virtually zero (.001 with a standard error of .056).

[22] We checked whether mixed teams do better/worse than homogeneous teams. The effect of gender is monotonic, with more women always being better.

C. Message Content: To better understand why teams without message costs do *not* out-perform individuals or teams with message costs, we turn to the messages sent by subjects in the team treatments. Our primary question is whether sending more messages is helpful or harmful and our primary measure of frequency of communication is the number of *relevant* messages sent per minute of work time. Messages are considered relevant if they relate to the task of solving the puzzle. To generate this measure we had two research assistants independently identify the relevant messages.[23] A fraction (10 percent) of messages were not relevant to the task at hand.[24] We generally ignore these messages since irrelevant messages are not expected to affect performance, but short discussions of the impact of irrelevant messages can be found below.

If frequency over the entire period is used, the number of messages sent by high-performing groups is biased downward because they solve the puzzle quickly and stop sending messages. We instead divide the number of messages by the time spent working to give the number of messages *per minute*. Time spent working is nine minutes for groups that did *not* solve the puzzle and the solution time for groups that did solve the puzzle.[25]

Table 3 breaks down the number of messages sent per person per minute by role (Leader/ Follower) and treatment (Team-No Cost/Team-Cost). The first figure in each cell is the mean number of relevant messages and the second is the mean number of off-task messages.

The number of relevant messages sent is highly sensitive to the role and treatment. It is unsurprising that leaders send far fewer relevant messages than followers. The most common types of relevant message either make specific suggestions for filling in the puzzle (e.g. "fill in

[23] The two counts were highly correlated with each other ($\rho = 0.99$). We use the average of the two counts to reduce the effect of coding errors.

[24] To familiarize subjects with the chat program, we asked them to send "Hello" as a message to their group at the beginning of Round 6. This let them see how to send a message and how messages from others would be displayed. "Hello" messages are by far the most common type of irrelevant message.

[25] Messages could not be sent once the problem was solved.

the bottom row") or point out mistakes (e.g. "C VII is wrong"). Since a leader can directly make and delete entries to the puzzle, there is little point in a leader sending these sorts of messages. Leaders send about the same number of off-task messages as followers.

Table 3: Frequency of Messages per Minute (Relevant/Off-Task)

	Team – No Cost	Team – Cost
Follower	1.62 / 0.16	0.47 / 0.04
Leader	0.57 / 0.15	0.15 / 0.05

The message cost treatment sharply reduces the number of relevant messages. The average subject in the Team-No Cost treatment sent 11.5 messages over the entire period. At one cent apiece, the cost of sending messages in the Team-Cost treatment was tiny compared to the marginal value of solving the puzzle. The simple model discussed in Section 4 predicts fewer messages in the Team-Cost treatment, but we were surprised by the magnitude of the effect given the low cost.

Table 4: Regression Analysis, Frequency of Messages

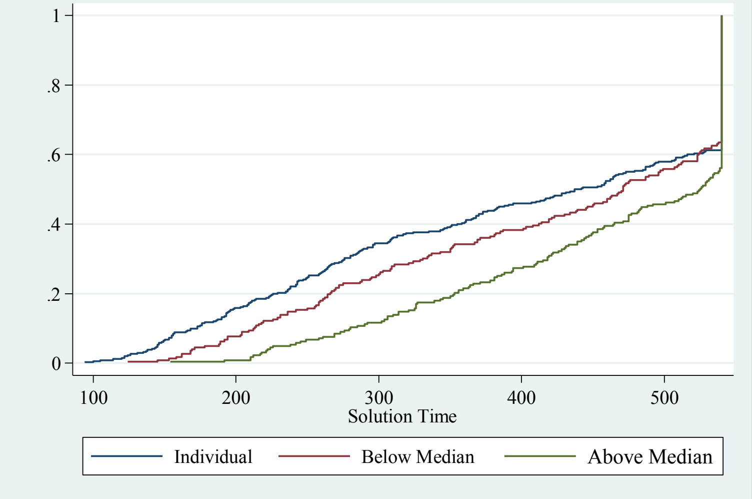
	Leader	Followers
Team–Cost	-0.581*** (0.149)	-1.253*** (0.117)
Time Bonus	-0.042 (0.149)	-0.55 (0.108)
High Pay	-0.064 (0.143)	0.137 (0.112)
Stage 1 Correct, Leader	-0.056 (0.043)	-0.610** (0.027)
Stage 1 Correct, Follower(s)	0.008 (0.011)	0.210*** (0.030)
Female	0.033 (0.086)	-0.024 (0.065)

p<0.1; ** p<0.05; *** p<0.01, two-tailed tests

Table 4 presents regressions taking a closer look at the frequency of messages. All data from the team treatments are used and separate regressions are run for leaders and followers. In both regressions the dependent variable is the number of relevant messages per minute sent by an individual. As independent variables, both regressions include a dummy for *Team-Cost*, dummies for the pay variations (time bonus and low vs. high pay), the number of puzzles solved in Stage 1 by the leader, a dummy for gender, and period dummies. The period dummies are not reported to save space. In the leader regression, we control for the number of puzzles solved in Stage 1 by all followers in the group, while in the follower regression we control for Stage 1 performance by the specific follower who generated the observation. A tobit model is used in both regressions since a number of subjects send no messages. The standard errors are corrected for clustering at the session level.

For both leaders and followers, the *Team-Cost* treatment has a strong negative effect on the number of messages sent. For followers, the number of messages sent is increasing in own Stage 1 performance and decreasing in their leader's Stage 1 performance. These results make sense: Low-ability leaders tend to need more help with the puzzles and high-ability followers tend to believe they have insights worth sharing. There are no significant gender effects.

Figure 6: Frequency of Talk and CDF of Solution Time



Underlying H1 and H3 is a basic assumption that getting suggestions from followers is helpful for a leader. The relatively poor performance of teams suggests that this is *not* the case. Figure 6 illustrates the relationship between the frequency of communication by followers and performance. For each treatment, we divide groups into sets according to whether they are below or (weakly) above the median number of relevant messages per minute sent by followers. Data from the *Individual* treatment is included as a point of reference. There is a large gap, both in terms of speed and likelihood of solving the puzzles, between those groups with a relatively low frequency of messages from followers and those with a relatively high frequency. Getting more messages from the followers seems to be *harmful* rather than helpful for the leader. This is rather surprising, since the only tangible cost of receiving a message is the time spent reading the message. Given that the average message is a mere 17 characters long (most messages are short, giving specific suggestions about filling in the puzzle), reading time should be minimal. If messages were typically beneficial, the help with solving the puzzle should more than compensate for the slight time spent reading them. This does not appear to be the case.

The effect shown in Figure 6 could reflect an indirect effect of leader ability, since this positively affects performance and more messages are sent to low-ability leaders. The regressions reported in Table 5 examine the effect of the frequency of followers' messages controlling for the ability of the leader.

Table 5: Effects of Follower Messages

Instrumental Variables	No	Yes
Correct Solution	-0.107** (0.053)	-0.172** (0.081)
Fast Solution	-0.310*** (0.085)	-0.265** (0.133)

$p<0.1$; ** $p<0.05$; *** $p<0.01$, two-tailed tests

We modify Models 2a and 2b from Table 2 to control for the number of relevant messages sent per minute, summing over the three followers. Recall that these regressions control for Stage-1 performance by both the leader and followers. The dataset is limited to observations from the team treatments and the dummy for the Individual treatment is dropped. Parameter estimates for the number of messages per minute by the three followers are reported in the first column of Table 5.[26] Increasing the number of messages sent by the three followers significantly decreases the likelihood of either a correct or fast solution. The effect is large as one additional message per minute decreases the probability of a correct solution by 4.0 percentage points and the probability of a fast solution by 5.5 percentage points.

Endogeneity could plausibly affect the results reported in the first column of Table 5 through two channels. First, the negative effects of follower messages on performance could reflect reverse causality if groups that are struggling with the nonograms send more messages per minute. The data suggest that this is not the case. Table 6 reports, by period, the number of relevant messages per minute sent by followers and the likelihood of solving the puzzles. The table is sorted by the percentage of correct solutions. As the puzzles get easier (more correct solutions), the number of follower messages increases rather than decreasing.

Table 6: Puzzle Difficulty and Message Frequency

Period	Follower Relevant Messages (per Minute)	% Correct
8	2.96	52.70%
6	2.94	53.50%
7	3.25	57.10%
9	3.61	65.90%
10	3.41	80.20%

[26] To save space we only report results for the variable of interest; full regression results are available upon request.

Second, increased follower messages could reflect attempts to help struggling leaders. Followers send significantly more messages to leaders who did poorly in Stage 1, plausibly reflecting a response to leaders who need help. This doesn't seem to explain the results in the first column of Table 5, since the regressions control for leader performance in Stage 1.

To directly address the possibility that endogeneity drives the results reported in this first column, we use the number of relevant messages per minute sent by the followers in *other* periods as an instrument for the number of relevant messages per minute sent by followers in the *current* period. This instrument is a good predictor of the current number of relevant messages per minute but should not be correlated with factors such as the leader's ability or the difficulty of the problem. The second column of Table 5 reports the results of the IV regressions. In all cases the estimates are similar to those in the original regressions and are statistically significant.

Conclusion 5: The performance of teams is decreasing in the number of messages sent by followers. The data provides no support for H3.

Going deeper into why increasing the number of messages harms performance, we examined the messages for what types of specific suggestions they contained. Specific suggestions are cases where a follower suggested either filling in a specific cell or unmarking a specific cell. We instructed coders to record the number of correct and incorrect suggestions in each message. A suggestion was correct if it called for filling in a cell that should have been marked in the correct solution or unmarking a cell that should *not* have been marked in the correct solution. Incorrect suggestions are defined in an analogous fashion. Many messages contained multiple suggestions, so a single message might count as multiple correct or incorrect suggestions. Specific suggestions are not the only type of relevant message sent by followers, but they are by far the most frequent type, relatively unambiguous to code, and obviously germane for solving the puzzles.

Good advice is common. Followers, as a group, averaged 5.08 correct suggestions per minute in *Team–No Cost* and 2.65 correct suggestions per minute in *Team-Cost*. Bad advice is rare, with averages of 0.49 and 0.23 wrong suggestions per minute in *Team–No Cost* and *Team-Cost* respectively. Given the prevalence of good advice, it is surprising that the effect of teams is not positive. Two things drive the weak performance of teams. First, even good advice doesn't help much. As a simple way of seeing this, divide the population by whether the followers are above the median number of correct suggestions per minute (by treatment, *Team-Cost* or *Team-No Cost*). The fraction of correct solutions is lower for groups with *more* correct suggestions (56 percent to 68 percent). Second, the effect of incorrect comments is large relative to the effect of correct ones. If we divide the population by whether followers exceed the median rate of wrong suggestions (by treatment, *Team-Cost* or *Team-No Cost*), the rate of correct solutions drops from 75 percent to 49 percent for the groups with *more* wrong suggestions.

Table 7: Parameter Estimates of the Effects of Suggestions

Dependent Variable	Correct Solution	Fast Solution (Solution Time < 270 seconds)
Relevant Messages per Minute	-0.066 (0.058)	-0.228** (0.109)
Correct Suggestions per Minute	0.006 (0.024)	-0.033 (0.060)
Wrong Suggestions per Minute	-0.761*** (0.212)	-0.956*** (0.275)

* p<0.1; ** p<0.05; *** p<0.01, two-tailed tests

Table 7 makes the same point more formally through regression analysis. Our starting point is Models 2a and 2b from Table 3. Recall that these include controls for the treatments, individual ability, and gender. As before, we restrict the sample to data from the team treatments and add a control for the number of relevant messages sent per minute by the three followers. We then add two new controls for the number of correct and wrong suggestions made by followers per minute. Table 6 reports the coefficients for the variables that relate to the frequency of messaging (the full output is available upon request).

The results shown are even more extreme than suggested by our informal analysis. Correct suggestions have little impact, while wrong suggestions account for much of the negative effect of having a team. The marginal effect of just a single wrong suggestion is large, reducing the likelihood of a correct (fast) solution by 29 (15) percentage points.

The preceding results contrast with the findings of Cooper and Kagel (2005, 2009, 2015) in a way that suggests why teams do very well in their games and poorly in our environment. When Cooper and Kagel look at dialogues between teammates, they find that the truth wins: good suggestions are almost always adopted and have an enormous positive impact while bad suggestions have a minimal effect. This is *not* the case in our experiments. Rather, good suggestions are having a minimal effect while wrong suggestions are catastrophic. It is a question for future research why the truth wins in one situation and not the other. This could be due to differences in teams structure – we use a leader follower structure while Cooper and Kagel requires teams to reach a unanimous agreement – or may flow from differences in the underlying problem – our problems require teams to have a long series of small insight while the game played in Cooper and Kagel hinges on teams having a single large insight.

This leaves us with a final question: if the effect of follower messages is strongly negative, and the *Team–Cost* treatment dramatically reduces the number of messages, why isn't the positive effect of the *Team–Cost* treatment larger? The detailed analysis points to a straightforward explanation: comparing the effects of follower messages across the two treatments is like comparing apples to oranges. The monetary cost of sending messages in the *Team-Cost* treatment does not depend on their content, but is per message. Sensibly, followers respond by packing more suggestions into each message. The number of specific suggestions per follower message climbs from 3.14 in *Team-No Cost* to 5.94 in *Team-Cost*, an 89 percent increase.[27] Likewise, the number of wrong suggestions per follower message climbs from 0.29 to 0.51. If the content of the messages changes with message costs, we would not expect messages to have the same effect in both treatments.

In line with this, we re-ran the regressions in the left column of Table 5 with an interaction between a dummy for *Team-Cost* and the number of relevant messages per minute sent by followers. This parameter captures the difference between the marginal effect of messages in *Team-No Cost* and *Team-Cost*. In both cases, the parameter is negative and statistically significant.[28] The negative effect of a message is roughly doubled in the *Team-Cost* treatment. Imposing a message cost gets followers to send fewer messages, which helps, but followers partially undo this positive effect by packing more into each message.

Our discussion has focused on messages sent by followers. The negative relationship between the number of messages sent by leaders and performance is even stronger than the relationship for followers, although there is the clear issue of endogeneity. The effect of leader messages is inherently less interesting than that of follower messages, since the benefit of having a team should, in theory, come through the flow of insights from the followers to the leader.

6. Concluding Remarks

[27] The difference is significant at the 1% level, based on a Wilcoxon rank-sum test on session averages.

[28] The parameter estimates are -0.187 and -0.209 for correct solutions and fast solutions respectively, with standard errors of 0.108 and 0.098.

There is a large experimental literature in economics that finds that groups are better at intellective tasks than individuals are, in the spirit that "two (or more) heads are better than one." But it may not be the case that teams will *always* out-perform individuals; the latter tend to arrive at decisions more quickly than groups (consider departmental meetings) and these decisions may well be better. Nevertheless, there is little or no previous incentivized experimental evidence that individuals are more effective than (or even as effective as) groups in a cognitive task.

Improved performance with teams is typically is driven by communication amongst the members of the group. In nearly all of this research, communication has no monetary cost. We have participants work on a difficult puzzle, either individually or in groups of four. Groups can communicate internally via unrestricted free-form chat. Individuals solve the puzzles more quickly than groups do, suggesting that messages from the followers interfere with the solution process. To alleviate this apparent congestion, we conduct sessions with a very small cost for sending a message. Imposing this cost dramatically reduces the number of messages sent and leads to a substantially improved speed of solution.

We also find an interesting gender effect: even though males out-perform females on the individual task, group performance improves monotonically with the number of females in the group and the magnitude of the improvement is considerable. This is consistent with previous literature suggesting that women have a higher degree of "social intelligence" and tend to work better in groups than males do.

The literature contains many examples where teams outperform individuals, but our study provides a cautionary note regarding group performance relative to individual performance, especially with respect to more communication always being better than less. In cases where too many messages are getting sent (i.e. messages have a negative externality), it may well be effective to limit communication by imposing a cost. We suggest that this principle might apply to environments with possible congestion effects, whereby imposing a very small cost for sending a written or verbal message may rein in counter-productive communication. Yet even here there is friction, as subjects appear to be good at gaming the system, getting around the message costs by cramming more suggestions into each message.

Further research is needed to determine the robustness of our findings in a variety of settings (i.e. different team structures, different problem, and different communication protocols) as well as identifying better methods of limiting the harmful aspects of communication while retaining the positive features.

References

Andreoni, J. (2014), "Trust, reciprocity, and contract enforcement: experiments on satisfaction guaranteed," Mimeo

Apesteguia, J., G. Azmat, and N. Iriberri (2012), "The Impact of Gender Composition on Team Performance and Decision Making: Evidence from the Field," *Management Science*, **58**, 78-93.

Ben-Ner, A., L. Putterman, and T. Ren (2011), "Lavish Returns on Cheap Talk: Non-binding Communication in a Trust Experiment, *Journal of Socio-Economics*, **40**, 1-13.

Blenko, M., M. Mankins, and P. Rogers (2009), *Decide & Deliver: 5 Steps to Breakthrough Performance in your Organization*, Bain and Company.

Blume, A., P. Kriss, and R. Weber (2014), "Pre-play Communication with Forgone Costly Messages, Mimeo.

Blume, A. and A. Ortmann (2007), "The effects of pre-play communication: Experimental evidence from games with Pareto-ranked equilibria," *Journal of Economic Theory*, **132**, 274-290.

Brandts, J., G. Charness, and M. Ellman (2015), "Let's Talk: How Communication Affects Contract Design," *Journal of the European Economic Association*, forthcoming.

Brandts, J. and D. Cooper (2007), "It's What You Say, Not What You Pay: An Experimental Study of Manager-Employee Relationships in Overcoming Coordination Failure," *Journal of the European Economic Association*, **5**, 1223-1268.

Brandts, J., D. Cooper, and R. Weber (2015), "Legitimacy, Communication, and Leadership in the Turnaround Game," *Management Science*, forthcoming.

Bray, R., N. Kerr, and R. Atkin (1978), "Effects of group size, problem difficulty, and sex on group performance and member reactions," *Journal of Personality and Social Psychology*, **36**, 1224-1240.

Casari, M., C. Jackson, and J. Zhang (2012), "When Do Groups Perform Better than Individuals? A Company Takeover Experiment," Mimeo.

Charness, G. (2000), "Self-serving Cheap Talk and Credibility: A Test of Aumann's Conjecture," *Games and Economic Behavior*, **33**, 177-194.

Charness, G. and M. Dufwenberg (2006), "Promises and Partnership," *Econometrica*, **74**, 1579-1601.

Charness, G. and M. Dufwenberg (2010), "Bare Promises," *Economics Letters*, **107**, 281-283.

Charness, G. and M. Dufwenberg (2011), "Participation," *American Economic Review*, **101**, 1211-1237.

Charness, G., E. Karni, and D. Levin (2007), "Individual and Group Decision Making under Risk: An Experimental Study of Bayesian Updating and Violations of First-order Stochastic Dominance," *Journal of Risk and Uncertainty*, **35**, 129-148.

Charness, G., E. Karni, and D. Levin (2010), "On The Conjunction Fallacy in Probability Judgment: New Experimental Evidence Regarding Linda," *Games and Economic Behavior*, **68**, 551-556.

Charness, G. and A. Rustichini (2011), "Gender Differences in Cooperation with Group Membership" *Games and Economic Behavior*, **72**, 77-85.

Charness, G. and M. Sutter (2012), "Groups make Better Self-Interested Decisions," *Journal of Economic Perspectives*, **26**, 157-176)

Cooper, D. and J. Kagel (2005), "Are Two Heads Better than One? Team versus Individual Play in Signaling Games," *American Economic Review*, **95**, 477-509.

Cooper, D. and J. Kagel (2009), "Equilibrium Selection in Signaling Games with Teams: Forward Induction or Faster Adaptive Learning?" *Research in Economics*, **63(4)**, 216-24.

Cooper, D. and J. Kagel (2015), " A Failure to Communicate: An Experimental Investigation of the Effects of Advice on Strategic Play," Mimeo.

Cooper, D., and K.-U. Kühn (2014), "Communication, Renegotiation, and the Scope for Collusion," *American Economic Journal: Microeconomics*, **6(2)**, 247-78.

Cooper, D. and M. Sutter (2014), "Endogenous Role Assignment and Team Performance," Mimeo.

Cooper, R., D. DeJong, R. Forsythe, and T. Ross (1992), "Communication in Coordination Games, *Quarterly Journal of Economics*, **107**, 739-771.

Dawes, R, J. McTavish, and H. Shaklee (1977), "Behavior, communication and assumptions about other people's behavior in a commons dilemma situation," *Journal of Personality and Social Psychology*, **35**, 1-11.

Fenwick, Graham, D., and Derrick J. Neal (2001), "Effect of gender composition on group performance," *Gender, Work and Organization*, **8(2)**, 205-225.

Fischbacher, U. (2007), "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, **10**, 171-178.

Greiner, B. (2004), "The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics," Mimeo.

Hackman, J. and N. Vidmar (1970), "Effects of Size and Task Type on Group Performance and Member Reactions," *Sociometry*, **33**, 37-54.

Isopi, A., D. Nosenzo, and C. Starmer, "Does consultation improve decision-making?", *Theory and Decision*, **77(3)**, 377-388.

Hoogendoorn, S., H. Oosterbeek, and M. van Praag (2013), “The Impact of Gender Diversity on the Performance of Business Teams: Evidence from a Field Experiment,” *Management Science*, **59**, 1514-1528.

Kochan, Thomas, Katerina Bezrukova, Robin Ely, Susan Jackson, Aparna Joshi, Karen Jehn, Jonathan Leonard, David Levine, and David Thomas (2003),The effects of diversity on business performance: Report of the diversity research network,” *Human Resource Management,* **42(1)**, 3-21.

Kugler, T., E. Kausel, and M. Kocher (2012). “Interactive decision making in groups. Are groups more rational than individuals?,” *Wiley Interdisciplinary Reviews: Cognitive Science,* **3***,* 471-482.

Laughlin, P., R. Bonner, and A. Miner (2002), “Groups perform better than the best individuals on letters-to-numbers problems,” *Organizational Behavior and Human Decision Processes*, **88**, 605–620.

Lorge, I. and H. Solomon (1955), “Two Models of Group Behavior in the Solution of Eureka-Type Problems,” *Psychometrika*, **20**, I39-148.

Oprea, R., G. Charness, and D. Friedman (2014), “Continuous Time and Communication in a Public-goods Experiment,” *Journal of Economic Behavior and Organization*, **108**, 212-223.

Shaw, M. (1932), “A comparison of individuals and small groups in the rational solution of complex problems,” *American Journal of Psychology*, **44**, 491-504.

Tindale, R., T. Kameda, and V. Hinsz (2003), “Group decision making: Review and integration,” In M. A. Hogg & J. Cooper (Eds.), *Sage Handbook of Social Psychology*, 381- 403. London: Sage.

Wilson, A. (2014), “Costly Communication in Groups: Theory and an Experiment,” Mimeo.

Woolley, A., C. Chabris, A. Pentland, N. Hashmi, and T. Malone (2010), “Evidence for a Collective Intelligence Factor in the Performance of Human Groups,” *Science*, **330 (6004)**, 686-688.

Appendix A: A Simple Model of a Follower's Decision to Send a Message

Consider the problem facing a follower within the team treatment without a time bonus facing the decision of whether or not to send a message.[29] The puzzle is currently in a state of partial completion. Assume that all subjects are thinking about the puzzle continuously and insights arrive over time via a Poisson process. When an insight arrives, what determines whether a follower develops and communicates his insight?

Let Δ_M be the increased probability of solving the puzzle if the message is sent and let π be the prize per person from solving the puzzle. Given that individuals may mis-estimate the impact of their messages, let $E(\Delta_M)$ be an individual's perceived value of Δ_M. For an overconfident individual, $E(\Delta_M) > \Delta_M$. It is possible (indeed, likely) that $E(\Delta_M) > 0 > \Delta_M$ for some messages. In other words, there are cases where an over-confident individual thinks they are sending a helpful message when their message is actually harmful (e.g. incorrect advice or irrelevant chatter that distracts other group members).

The cost of sending a message is broken into two components. The monetary cost of sending a message is denoted by c_m. This is the 1¢ per message cost from the *Team – Cost* treatment. Sending a message also has a non-pecuniary cost, c_e. This captures several types of costs. First there is the effort cost involved in thinking about an insight to the point it becomes useful. Ideas don't arrive fully formed and it takes some effort and thought to bring them to fruition. There is also an effort cost involved with typing in a message. Finally, sending a message may generate utility or disutility independent from the effort spent on producing the message. Some people like the sound of their own voice or enjoy coming up with a (hopefully)

[29] The problem facing a leader is similar, but not identical. Followers are typically passing suggestions on to leaders. Leaders are usually either asking for suggestions or helping develop ideas about how to solve the puzzle. They don't need to make suggestions to themselves. Assuming there is no time bonus is done to simplify the model and is not central to any of our conclusions.

funny message. They derive positive benefits (i.e. negative costs) from sending a message. Likewise, a shy individual may experience disutility from sending a message. This is incorporated into c_e as a cost.

When an insight arrives, a follower develops and communicates his insight if the perceived benefit is greater than the cost. In other words, a message is sent if the inequality shown as (eq. 1) holds.

$$\mathrm{E}(\Delta\mathrm{M})\pi > c_E + c_m \quad \textbf{(eq. 1)}$$

For many parameter values, this model mirrors a standard public-goods (or joint-production) problem. The benefits on the left side of (eq. 1) accrue to the person sending the message, but the benefits accruing to other group members are absent. Assume $c_E + c_m > 0$. If the follower is well calibrated ($\mathrm{E}(\Delta\mathrm{M}) = \Delta\mathrm{M}$) and the message is valuable ($\Delta\mathrm{M} > 0$), sending a message generates a positive externality and a self-regarding follower sends too few messages from a social point of view. The problem is mitigated with social preferences for the usual reasons.

However, it is easy to devise plausible cases where sending a message generates a negative externality. One obvious case occurs when a follower's insight is, in reality, mistaken. This implies $\mathrm{E}(\Delta_\mathrm{M}) > 0 > \Delta_\mathrm{M}$. Too many messages are sent from both an individual and a social point of view. Similar logic applies to the case where a follower fails to account for the congestion caused by his message. Due to the resulting negative externality, too many messages are sent. Another possibility is that a follower is correctly calibrated and understands his messages do not help solve the puzzle but puts high intrinsic value on sending messages, such that $0 > \mathrm{E}(\Delta\mathrm{M})\pi = \Delta\mathrm{M}\pi > c_E + c_m$. This sort of follower, while sending the optimal number of messages from an individual point of view, sends too many messages from a social point of view.

Adding a message cost, as occurs in the *Team–Cost* treatment, increases c_m. If there is a positive externality associated with sending messages, adding a message cost exacerbates the problem and reduces performance. If there is a negative externality, it is helpful to limit the number of messages sent and performance improves.

Appendix B: Alternative Definitions of Fast Solution

. Table B.1 addresses how robust our results are to varying the definition of a Fast Solution. This problem by systematically varying the cutoff used to define a fast solution. The cutoff used for solving the problem is listed in the left column. It starts with a low cutoff of 200 seconds, a speed achieved in less than 10% of the sample, and then increases the cutoff in increments of 50 seconds. We also include cutoffs of 270 seconds (half the available time, the cutoff for a fast solution used elsewhere in this paper) and 540 seconds (i.e. did the individual/group solve the problem in the available time). For each definition of a fast solution we use the same specification as in Models 1a and 1b. For a cutoff of 270 seconds the regression is identical to Model 1b and for a cutoff of 540 seconds it is identical to Model 1a. For each cutoff we report the parameter estimate and standard error (corrected for clustering) for Individual and Team-Cost.

Table B.1: Changing the Definition of a Fast Solution

Seconds	Individual	Team-Cost
200	$.736^{***}$ (.201)	.398 (.247)
250	$.573^{**}$ (.221)	$.418^{**}$ (.212)
270	$.462^{**}$ (.192)	$.448^{**}$ (.200)
300	$.418^{**}$ (.191)	$.444^{**}$ (.218)
350	$.344^{*}$ (.207)	$.497^{**}$ (.215)
400	.203 (.200)	$.365^{*}$ (.204)
450	-.015 (.192)	.177 (.189)
500	-.084 (.176)	.116 (.194)
540	-.266 (.168)	.073 (.186)

* p<0.1; ** p<0.05; *** p<0.01

Appendix C: Instructions (Team – Cost, Low Pay, No Time Bonus)

Experiment Instructions

Welcome

This is an experiment in the economics of decision-making. Several research institutions have provided funds for this research. You will be paid for your participation in the experiment. The exact amount you will be paid will depend on your and/or others' decisions. Your payment will consist of the amount you accumulate plus a $10 show-up fee. You will be paid privately via check at the conclusion of the experiment.

If you have a question during the experiment, raise your hand and an experimenter will assist you. Please do not talk, exclaim, or try to communicate with other participants during the experiment. Please put away all outside materials (such as books, bags, notebooks, cellphones) before starting the experiment. Participants violating the rules will be asked to leave the experiment and will not be paid.

Basic Structure

Today you will solve a series of puzzles called nonograms. First you will learn the basic rules of the puzzle and then you will try your hand at solving some nonograms individually. You will be paid $0.50 for each nonogram you solve within the time limit in the individual stage.

After that, we will form groups of four participants and you will solve more difficult nonograms with your group. You will paid $3 for each nonogram your group solves within the time limit in the group stage. You will be able to consult with other members of your group using a chat box, but you will be charged a small fee for each chat message.

What is a Nonogram?

A nonogram is logic puzzle. The rules are simple. You have a grid of squares, each of which must either be marked with an X or left blank. Beside each row of the grid are listed the lengths of each "run" of continuous X's in that row, in order. Above each column are listed the lengths of each "run" of continuous X's in that column, in order.

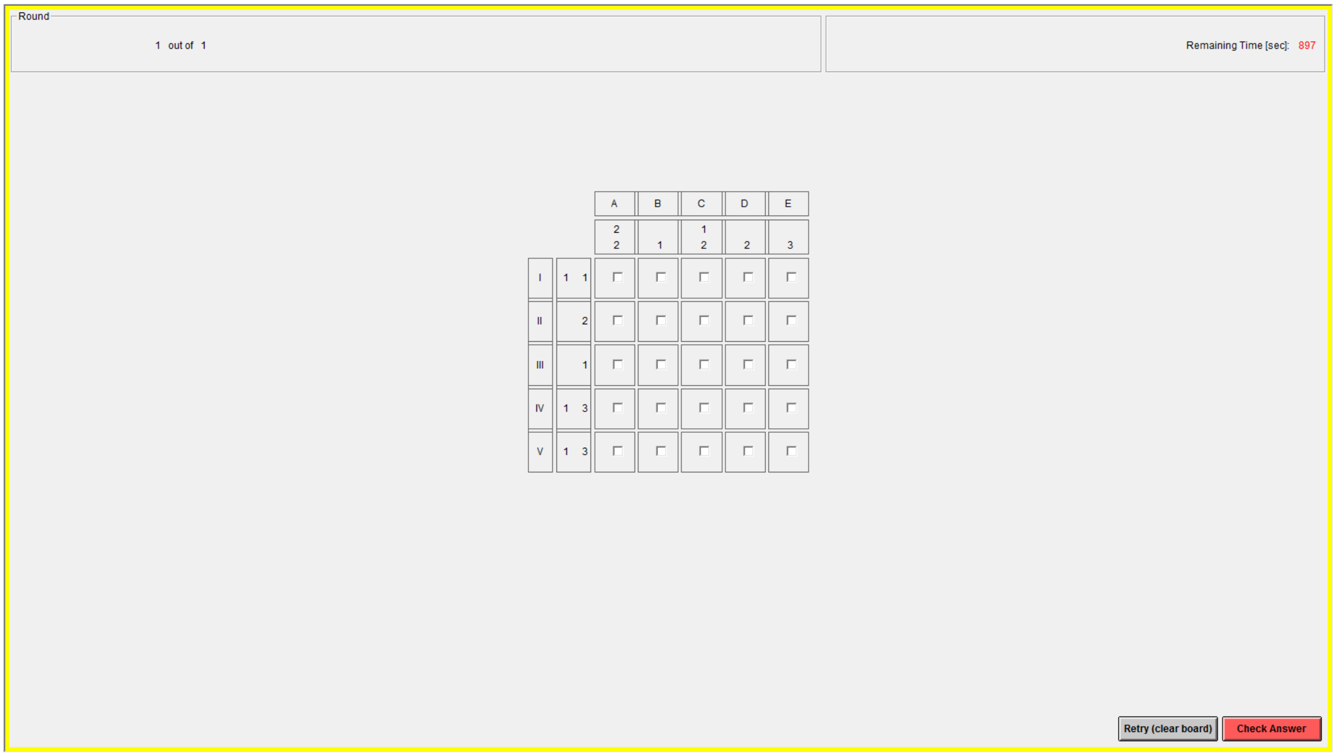
Your aim is to find and mark all the X's. Click on a square to mark it with an X and click on it again to unmark it. If you would like to clear all the X's from the grid, click "Retry (clear board)". If you have finished, click "Check Answer".

Below is an example screenshot of a 5x5 nonogram (5 rows, 5 columns). For reference, we have labeled the rows with roman numerals I – V and the columns with letters A – E.

Note that row II is labeled with a "2". This means that there are two X's in a row, with no spaces between them, somewhere in that row.

Row IV is labeled with a "1" and a "3". This means that in that row, there is one X and then three X's in a row, separated by one or more blank spaces.

Finally, look at column C, which is labeled with a "1" above a "2". This means that there is one X and then two X's in a row, separated by one or more blank spaces.

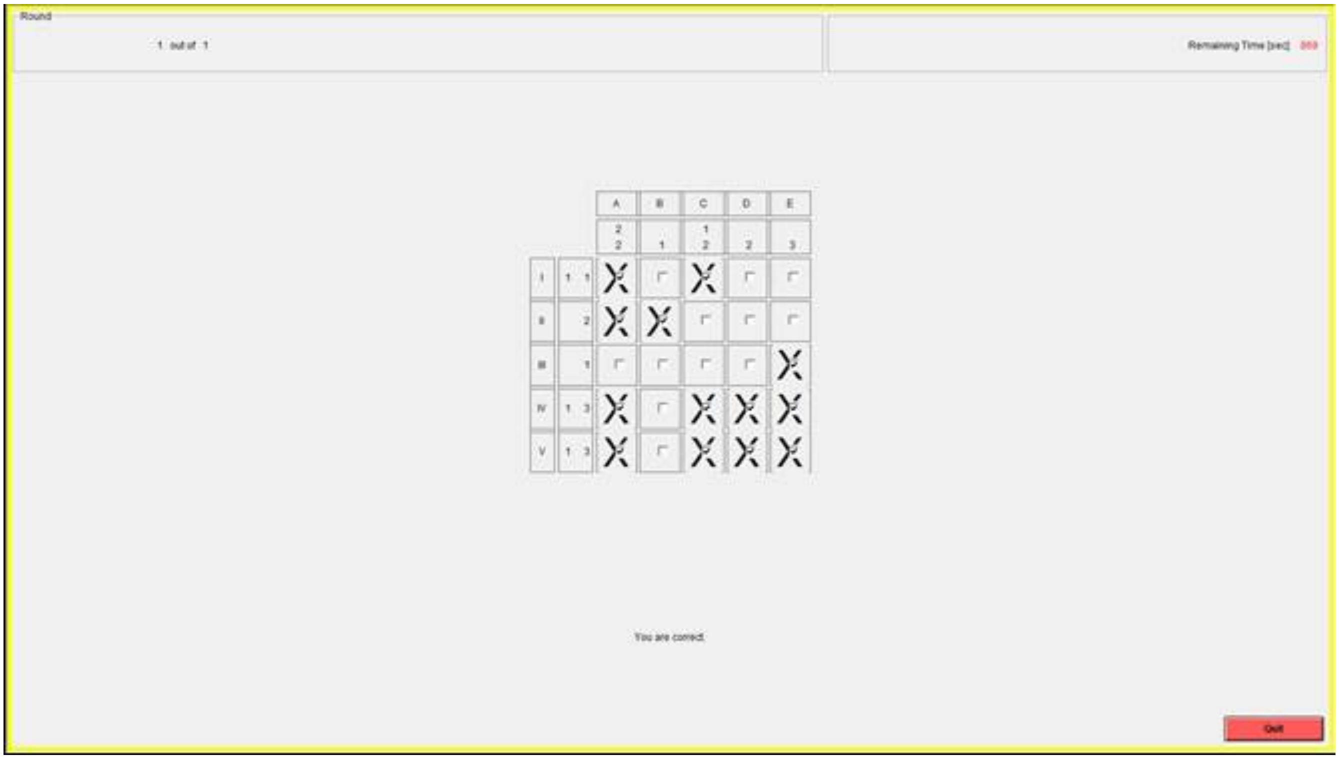


1 out of 1

Remaining Time [sec]: 897

		A	B	C	D	E
		2 2	1	1 2	2	3
I	1 1	☐	☐	☐	☐	☐
II	2	☐	☐	☐	☐	☐
III	1	☐	☐	☐	☐	☐
IV	1 3	☐	☐	☐	☐	☐
V	1 3	☐	☐	☐	☐	☐

Below is a screenshot of what this puzzle looks like when it is correctly completed.



Round

1 out of 1

Remaining Time [sec] 869

		A	B	C	D	E
		2 2	1	1 2	2	3
I	1 1	X	□	X	□	□
II	2	X	X	□	□	□
III	1	□	□	□	□	X
IV	1 3	X	□	X	X	X
V	1 3	X	□	X	X	X

You are correct.

Before learning more details, we will pause for questions.

Individual Stage

You will have 95 seconds to complete each of the five nonograms in the individual stage. Each puzzle has 5 rows and 5 columns (5x5), and if you complete it before the time is up, you will earn 50 cents.

Group Stage

There are five rounds in the group stage. In each round, everyone will be randomly assigned into groups of four. You are ***not*** with the same group for all five rounds; in each round you will be assigned to a ***different*** group of people. The groups are anonymous—you will not know which of the other people in the room are in your group. In each round, your group will work together to solve a 10x10 nonogram with a time limit of 540 seconds (9 minutes).

Although everyone in the same group can see the puzzle, only one person, called the Leader, will be able to mark or unmark the squares on the board. Other than this, there is no distinction between the group's Leader and the rest of the people, who are called Followers. You will be in the same role (Leader or Follower) for all five rounds.

Players in the same group may communicate with and advise each other using a chat box. This will be available until you have solved the problem or the time limit is reached. Chat messages will be displayed to all other members of the same group. Only people in your group will be able to see your messages. Each person in the group has a chat id (1 – 4) so you can identify who is sending messages. The chat ids are randomly generated in each round, so your chat id will change from round to round.

The chat box works much like an IM program. You have a box to type messages. When you hit the enter key, all four members of the group see your message. When using the chat box, we ask that you follow two simple rules: (1) Do not send any messages that would allow another group member to identify you and (2) please avoid offensive language.

If your group completes a puzzle within the time limit, each person in the group will earn $3. If you do not solve the puzzle, each person will earn $0. Every chat message that you send will cost you $0.01, regardless of whether or not your group completes the nonogram.

Example 1: If your group completes the nonogram and you sent 10 chat messages, you will earn $2.90 = $3.00 - $0.10 for the round.

Example 2: If your group does not complete the nonogram in the 9 minutes and you sent 20 chat messages, you will *lose* $0.20 for the round.

Your earnings for the session are calculated by adding your earnings from the five individual rounds to your earnings from the five group rounds, as well as your $10 show-up fee.