TI 2018-070/VII
Tinbergen Institute Discussion Paper

Committees of Experts in the Lab

Sander Renes[1]
Bauke (B.) Visser[1]

[1] Erasmus University Rotterdam

Committees of Experts in the Lab

Sander Renes and Bauke Visser*

August 30, 2018

Abstract

Theory predicts that committees of experts may take decisions that look good but are bad and that they show a united front to impress evaluators. Although evaluators see through this behavior, committees persist in it only to avoid worse assessments. We investigate this theory in the lab, using treatments with and without reputation concerns and with and without cheap-talk communication with evaluators. We use the chat among committee members to learn about, *e.g.*, their beliefs about the determinants of evaluators' assessments. We find that a committee's desire to come across as well-informed causes it to garble the information on which evaluators can base their assessments. Evaluators see through this behavior, making their assessments less dependent on actual decisions and statements. With or without reputation concerns, for the majority of committees, words speak louder than costly decisions. Evaluators pick this up. Orthogonality tests show that evaluators use observable clues about ability quite efficiently but struggle to infer ability from infrequent statements. The absence of cheap talk as a means to influence assessments hurts decision making and reduces the overall accuracy of assessments. Evidence that united fronts are consciously formed is limited.

Keywords: committees, reputation concerns, assessments, cheap talk, united front, information garbling

JEL codes: C91, D71, D83, D84, L14

*Renes: Erasmus University Rotterdam, Tinbergen Institute, ERIM and SFB884 Political Economy of Reforms, srenes@ese.eur.nl. Visser: Erasmus University Rotterdam and Tinbergen Institute, bvisser@eur.nl. We thank Sebastian Fehrler, Chaim Fershtman, Lúis Santos-Pinto, Karl Schlag, and seminar audiences at Erasmus University Rotterdam, Middlesex University and at the universities of Mannheim, Konstanz, Lausanne, Milan and Vienna for helpful comments and discussions. Annikka Lemmens and Erik van Goudoever provided diligent research assistance. We gratefully acknowledge financial support by Netherlands Organisation for Scientific Research (NWO) through grant 400-09-338 and Erasmus University Rotterdam through CSTO grant 2014–54.

1 Introduction

Decision making committees are frequently used to bring together experts on a specific matter. Monetary policy committees decide on key interest rates, health care consensus panels on medical protocols and senior management teams on strategic matters, corporate or public. There is a growing literature that explains the behavior of members of such bodies, and of decision makers more generally, in terms of a specific form of reputation concerns – a concern to come across as well-informed.

Theory predicts that the desire to impress an evaluator may lead to decisions that look good, but are bad. A rational evaluator, however, is not fooled by the behavior of decision makers. She anticipates their interest in interfering with the conclusions she draws about their ability. Nevertheless, decision makers persist in their behavior, to avoid worse assessments.

Variations of this interaction between decision makers and an evaluator – which was described first by Holmström (1999) – have been used to explain, e.g., herd behavior, biased forecasts and advice, rash decision making giving way to conservatism, self-censorship in committee meetings and united fronts towards the outside world, and various undesired reactions to transparency imposed on committees.[1] More generally, and more positively, Fama (1980) argued that career concerns play an important role in explaining why the separation of ownership and control can be an efficient form of organization.[2] Others, including Dewatripont et al. (1999a,b), have argued that career concerns play an even more important role in the public sector than in the private sector.

Equilibria in these models require a high degree of strategic sophistication. Despite its importance in the theoretical literature, and the obvious relevance for real-world decision making and governance, little is known about the equilibrium relationship between decision makers and evaluators in practice. In part, this is caused by the lack of observability of key factors in the model. On the one hand, to establish whether a decision maker takes particular decisions that look good but are bad, one should know what the correct decision is. On the other hand, to measure the quality of the evaluator's assessment of a decision maker one should know his true characteristics. Neither is easily established by an outsider on the basis of

[1] For herd behavior, see Scharfstein and Stein (1990) and Ottaviani and Sørensen (2001); for biased forecasts and advice, see Ottaviani and Sørensen (2006a,b); for rash juniors and conservative seniors, see Prendergast and Stole (1996); for behavior in committees, see Visser and Swank (2007), Levy (2007) Swank and Visser (2013), Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2017).

[2] The above-mentioned paper by Holmström, first published in 1982, was the first attempt to understand under what conditions Fama's claim as to the efficient choices induced by career concerns was true.

observational data.[3] A second reason may be that the evaluator is modeled as a disinterested machine that dutifully applies Bayes' rule to equilibrium behavior of decision makers. Human evaluators, however, may struggle interpreting the actions of decision makers, especially of those with an interest in positive assessments. This could provided incentives for decision makers to distort their actions in manners unpredicted by theory.

To overcome these observability problems, and to replace the evaluating machine by humans, we designed a laboratory experiment. In this experiment, half of the subjects form two-member committees that take a binary decision under uncertainty, while the other half evaluates whether decision makers are well-informed. The set up closely follows the committee-decision model of Visser and Swank (2007; VS from now on). Roughly speaking, the experiment integrates a voting experiment for committee-member subjects with a subjective probability elicitation experiment for evaluator subjects. The probability assessments concern the ability level of the committee members and are based on their observable actions. To emulate a reputation concern, part of the payoffs of the committee members is determined by probabilities elicited from evaluators.

In VS, members receive a private signal about the state, deliberate and vote to take a binary decision on a project, all behind closed doors. Finally, evaluators observe the decision taken – but not the true state – and cheap-talk statements sent by committee members about anything that prevailed in the meeting. Members care both about taking the right, state-dependent decision and about their reputation for competence. This reputation is defined as the end-of-game probability that a member is of high ability according to the evaluator. The evaluator's role is to determine this probability. A key prediction of the model is the decision on the project in case of conflicting private signals. In the model, conflicting signals imply that the best decision is to reject the project *and* that at least one member is of low ability. The latter stems from the fact that a high-ability member receives a signal equal to the state (and thus equal to the signal of another high-ability member), while a low-ability member receives an uninformative signal. Thus, an evaluator who believes that a committee decides to implement the project after two positive signals and to reject it after two conflicting or two negative signals, rationally assigns a higher reputation after implementation than after rejection. This creates a dilemma if members receive conflicting signals and care about their reputation. From a project-value perspective, the project should be rejected; from a

[3]As a result, empirical work has focused either on intertemporal patterns of a manager's compensation that can be explained by career concerns or on industries where market-based incentives can be measured. See Hermalin and Weisbach (2017) for a review of that literature.

reputation perspective, the project should be implemented. The drop in reputation from rejection makes that, in equilibrium, reputation-concerned members distort the decision – they implement the project with some probability even in case of conflicting signals. A rational evaluator understands this inclination. As a result, the gain in reputation from implementation is smaller than what it would have been had members not been concerned with their reputation. The model also predicts that if committee members care about their reputation, cheap-talk communication with an evaluator contains no information about members' abilities. Indeed, VS argue that members form a united front to the outside world, regardless of what has occurred in the meeting. An evaluator thus rationally ignores any cheap-talk statement and bases her assessments on the decision on the project alone.

An advantage of studying a committee – rather than a single agent – is that conversations about what to vote and what statement to send to an evaluator form a natural part of the decision-making process. In the experiment, conversations were computer-mediated, thus creating a rich source of information on, *e.g.*, decision makers' beliefs concerning the relationship between their actions and assessments. Access to such information is particulary useful if observed behavior were to differ from equilibrium behavior.

The presence of reputation concerns in members' objective function is likely to complicate both committee members' choice of actions and evaluators' formation of beliefs, as it creates a strategic interaction between these beliefs and the choices of committee members. Similarly, we expect that if members can use both cheap talk and the binary decision, then evaluators would find it harder to assess members due to the greater amount of available information, while committees would find it harder to choose actions. The model yields a rich set of predictions about committee behavior and related assessments, both when members only care about the state-dependent decision payoff and when they also care about assessments, as well as in the presence and absence of cheap talk.

We designed the experiment to answer a number of questions. How do reputation concerns change the decisions that committees take and the accompanying cheap talk statements? Does the possibility of using cheap talk to communicate with their evaluators change the way decision makers attempt to influence assessments? Is their behavior a best-reply to evaluators' assessments? Does knowledge that committee members are reputation concerned change the way or the accuracy with which evaluators assess them? How good are evaluators at combining information about committee members' ability contained in costly decisions and cheap-talk statements?

The main treatment has two characteristics. First, a committee member (he) has a dual objective. The payoff he receives equals the sum of a decision payoff and the average assessments received from the evaluators (she). The decision payoff depends on whether the committee's decision, $Y = 1$ or $Y = 0$, matches the state.[4] Second, an evaluator – who is incentivized to submit her true assessment – observes two pieces of information before submitting her assessment of a committee member. The first piece is the decision a committee took. The second piece is the (cheap talk) statement of each committee member about his confidence in the decision taken. A member could choose one of five statements, ranging from 'Very Doubtful' to 'Very Confident.' This treatment is called *A-Stm*, to highlight that the payoffs of decision makers also depend on the Assessments they receive and that their decisions are accompanied by cheap talk Statements. Each of the other two treatments changes one of these characteristics. In the *NoA-Stm* treatment, the payoffs of a committee member are independent of assessments; they equal the decision payoff. Any strategic interaction between committees and evaluators is absent. Nevertheless, committee members do make statements about their confidence in their decision. Moreover, as in the *A-Stm* treatment, evaluators do assess members on the basis of the decisions taken and the statement made. We use a comparison of the outcomes of the *A-Stm* and *NoA-Stm* treatments to establish how committee members use decisions and statements to shape their assessments and, vice versa, how evaluators react to decisions and statements that are known to come from members who are reputation concerned. In the third treatment, the *A-NoStm* treatment, a member does not make statements about his confidence in the decision, but he still cares about assessments and the state-dependent decision payoffs. As a result, evaluators only observe the decision of the committee. We use a comparison of the outcomes of the *A-Stm* and *A-NoStm* treatments to establish any effect of the presence of cheap-talk statements on the assessments of evaluators and on the decisions that committees take.

In the experiment, like in VS, a committee member either receives an informative signal that matches the state - and is thus well-informed - or receives an uninformative signal that is unrelated to the state. Moreover, a member only knows that he is well-informed with a certain probability. We designed a novel scheme that builds on the traditional urn scheme to explain to subjects the relationship between the state and his ability on the one hand and the signal he receives on the other.

Our experimental results show, first, that evaluators are clearly aware whether

[4]In VS, $Y = 1$ stands for project implementation while $Y = 0$ denotes rejection; in the experiment, which is cast in an urns-and-balls framework, $Y = 1$ stands for Yellow and $Y = 0$ for Blue.

committee members have a strategic interest in obtaining positive assessments: assessments in the *A-Stm* treatment depend less on observed members' actions – decision and statement – than in the *NoA-Stm* treatment. We also find that evaluators shift their attention from decisions to statements when the latter are available to members as an instrument to shape assessments. Second, the conversations in the *A-Stm* treatment reveal that committee members pick up that their assessments depend on their statements, rather than on their decisions. This gives rise to conversations about what statement to make, rather than what decision to take, to influence assessments. As a result, and contrary to the theory, the availability of cheap-talk statements leads to a considerable reduction in the frequency with which the decision is distorted. Third, reputation concerns drastically inflate the confidence that committees express in the decision taken and make that among committee members the modal statement strategy stops revealing any information about members' private information and thus about their abilities. Finally, the statement strategies of a substantial majority of members remain informative about their abilities. Although a large minority of committees ends up using the same statements, only two committees do so after discussing the presumed benefits of showing a united front.

Observed behavior, of committee members and evaluators alike, deviates in some dimensions from equilibrium behavior. A key question is then whether evaluators best-reply to observed committee behavior. We use orthogonality tests to see whether evaluators make efficient use of observed committee behavior. We find that they use the available information quite efficiently in all treatments, but that they have difficulties in dealing with infrequent statements. Turning to committee members, we find that in the *A-Stm* treatment, committee members who revealed nothing about their private signals through their cheap-talk statements and who did not distort the decision on the project earned the highest payoffs. In the *A-NoStm* treatment, the highest-earning committee members were those that did distort the decision: their gain in assessment more than outweighed the expected loss on the project.

We conclude with an information-theoretic analysis. We use a formal measure of information, entropy, to determine the extent to which the information that the computer makes available about members' abilities – in the form of a pair of private signals – is garbled by committee members and is next picked up by evaluators. Entropy, being a cardinal measure of information, also allows us to make comparisons across treatments. The analysis shows that reputation concerns half the information that evaluators can glean from observed committee behavior. The drop is particu-

larly large if evaluators only observe the decision. This is because in this experiment, words speak louder than actions. They speak 5 to 15 times louder depending on the treatment.

There are a few other experiments that investigate how a concern with coming across as well-informed affects behavior. Berg et al. (2009) used their experiment to show that decision makers' commitment to a chosen, but erroneous course of action, is better explained by such reputation concerns than by a concern for consistency per se.

Like us, Meloso et al. (2017) aim to understand the interaction between a sender who cares about coming across as well-informed and an evaluator, by comparing behavior in treatments that vary in terms of the complexity of the interaction. Unlike us, they study a single sender who is exclusively interested in coming across as well-informed and can only use cheap talk to communicate with an evaluator who gives her assessment after observing the realized state. Moreover, they focus on the behavior of the sender by varying whether the assessments come from computerized evaluators of varying degrees of sophistication or from a human subject.

Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2017) study behavior of a committee of reputation-concerned decision makers, but their focus is different. They study the effect of secrecy and transparency of the decision-making process on the behavior of subjects and the quality of decisions taken.[5]

Others have found that theory may underestimate the amount of private information that senders reveal in the lab. This phenomenon has been called 'overcommunication,' but the focus has been on contexts in which senders can tell the truth or lie about a privately received signal.[6] However, claiming to be very confident, say, in the decision even though one's committee received conflicting signals is not a lie. In the experiment, as in VS, senders can use both cheap-talk statements and costly signals – the decision on the project – to influence receivers' behavior. Only costly decisions are predicted to be effective in doing so. We find that overcommunication in cheap talk means underutilization of the costly signal and thus a reduction in the distortion.

The rest of the paper is organized as follows. The next section presents the the-

[5]Other experiments, like Koch et al. (2009), Irlenbusch and Sliwka (2006) and Katok and Siemsen (2011), study subjects who want to come across as able in contexts in which ability together with effort determine observed performance.

[6]If theory predicts a sender to lie about his signal, but the subject in the lab truthfully reveals it, the subject is said to 'overcommunicate.' See Dickhaut et al. (1995) and Cai and Wang (2006). See also Goeree and Yariv (2011) and Fehrler and Hughes (2018) in a committee setting. Meloso et al. (2017) find both overcommunication and 'undercommunication' – senders misreporting their private information where theory predicts truthful revelation.

ory of VS, while the experimental design is described in section 3. Section 4 presents the findings and section 5 compares these findings with the theoretical predictions. In section 6, we present an information-theoretic analysis of the treatments. Section 7 concludes. Various robustness checks are presented in the Appendix. The instructions for the experiment can be found on our websites.[7]

2 A theory of decision making by career-concerned committees

The experimental design follows a simplified version of the model of VS: committees consist of two, rather than n, members; and members are homogenous, rather than heterogenous, in the weight they attach to their reputation. The latter simplification means that, at least in theory, there are no conflicts within the committee. The focus of the experiment is on the interaction between the committee and the evaluators. Thus, the two-member committee decides whether to implement a project, $Y = 1$, or reject it, $Y = 0$. Rejection yields a 'project payoff' equal to zero. The payoff of implementation is uncertain and state dependent. It equals $p+\mu$, where $\mu \in \{-h, h\}$ with $\Pr(\mu = h) = 1/2$. Thus, μ denotes both the state and the state-dependent part of the payoff. Ex ante, the expected value of implementation is $p < 0$.

At the start of the game, Nature determines both the state, μ, and the ability level of each member $i = 1, 2$, $a_i \in \{\underline{a}, \bar{a}\}$, with $\Pr(a_i = \bar{a}) = \pi$, where $\bar{a}$ stands for high ability, while $\underline{a}$ denotes low ability. Nature does not inform anyone about either μ or a_i. Next, each member receives a private signal $s_i \in \{s^g, s^b\}$ about the state μ. The quality of the signal member i receives depends on a_i. If i is highly able, he receives a high quality signal, $\Pr(s_i = s^g \mid \mu = h, a_i = \bar{a}) = \Pr(s_i = s^b \mid \mu = -h, a_i = \bar{a}) = 1$. If i is of low ability, he receives a low quality signal, $\Pr(s_i = s^g \mid \mu = h, a_i = \underline{a}) = \Pr(s_i = s^b \mid \mu = -h, a_i = \underline{a}) = 1/2$. Thus, the prior likelihood that a private signal matches the state is $(1 + \pi)/2 > 1/2$. In the deliberation stage that follows, each member sends a cheap talk message $m_i \in \{m^g, m^b\}$ to the other member. This message can be related to his private signal. Following the deliberation stage, members cast a vote on the project, $v_i \in \{v^1, v^0\}$, where $v_i = v^1$ denotes that i votes for $Y = 1$, and $v_i = v^0$ means a vote for $Y = 0$. $Y = 1$ requires both members to vote v^1. The decision taken by the committee is observed by the 'market.'

Finally, in the statement stage – a stage not to be confused with the deliberation

[7] https://personal.eur.nl/bvisser/ and http://sanderrenes.com.

stage – each committee member decides what cheap-talk statement ω_i to send to the market. This statement can be about anything that prevailed in the meeting. Let $\omega = (\omega_1, \omega_2)$. Next, the market determines the updated belief that a member is well-informed on the basis of Y and ω, $\hat{\pi}_i(Y, \omega) = \Pr(a_i = \bar{a}|Y, \omega)$. The objective function of a member equals $U_i(Y, \mu) = Y \cdot (p + \mu) + \lambda \hat{\pi}_i(Y, \omega)$, where $\lambda \geq 0$ denotes the weight members attach to the market's assessment. Parameter values are such that, from a project-value perspective, the project should be implemented iff $(s_1, s_2) = (s^g, s^g)$, and should be rejected in case of conflicting signals (and thus also if $(s_1, s_2) = (s^b, s^b)$).

A key feature of the model is that conflicting signals 'cancel each other out' in terms of expected project-value, $\mathbb{E}\left[\mu|s^g, s^b\right] = 0$ *and* are a sure sign that at least one member is uninformed. The first result follows from the fact that both members are equally likely to be of high ability. The second follows from the relationship between ability and signals received: two high ability members receive the same signal with probability one.

Models with deliberation and voting have many equilibria, and the one of VS is no exception. We focus on cheap talk deliberation strategies that are informative if they exist and voting strategies that are undominated.

For $\lambda = 0$, the equilibrium deliberation strategy is to truthfully reveal the private signal and the undominated voting strategy is to vote for implementation if both messages - signals, really - are positive, and to vote for rejection in the remaining cases. This voting strategy maximizes the expected project payoff. As statements have no payoff consequences, theory does not predict a statement strategy.

Suppose now that $\lambda > 0$ and that members cannot send cheap-talk statements to the evaluator. The evaluator's assessment is then based on Y. If members care little about assessments, the committee chooses $Y = 1$ if and only if both signals are positive. As a result, $\hat{\pi}(Y = 1) > \hat{\pi}(Y = 0)$, as the evaluator infers from $Y = 1$ that private signals are the same (and positive), whereas $Y = 0$ may mean that committee members received conflicting signals. If members care considerably about their assessments, they deviate from the voting strategy that maximizes expected project payoff if they receive conflicting signals. They do so if $\lambda > -p/\left[\hat{\pi}(Y = 1) - \hat{\pi}(Y = 0)\right]$. Define $\beta = \Pr(Y = 1|s_1 \neq s_2) \in [0, 1]$ and let $\hat{\pi}(Y; \beta)$ denote the equilibrium assessment if the committee chooses Y and distort the decision on Y with probability β conditional on conflicting signals. In equilibrium, β satisfies

$$p + \lambda \hat{\pi}(Y = 1; \beta) = \lambda \hat{\pi}(Y = 0; \beta). \tag{1}$$

Thus, if members have received conflicting signals, what a member gains in assess-

ment thanks to implementation offsets the expected loss due to the distorted decision on the project. As $p < 0$, $\hat{\pi}(Y = 1; \beta) > \hat{\pi}(Y = 0; \beta)$ holds in equilibrium. This requires that, conditional on conflicting signals, committees are less likely to choose $Y = 1$ than $Y = 0$, $\beta < 1/2$. Thus, implementation remains a sign of members being well-informed in equilibrium, albeit a weaker one than in the absence of a concern with assessments. From (1) it also follows that with two positive signals, committee members prefer $Y = 1$, whereas with two negative signals they prefer $Y = 0$. Finally, for any λ, members share their private information in the deliberation stage.

Now assume that the market also observes the pair of cheap-talk statements ω when assessing committee members. VS establish that, for a given Y, any ω that the committee uses in equilibrium leads to the same assessment. Had this not been the case, members would always choose the statement pair with the higher assessment. This has two implications. First, the market bases its assessment exclusively on the decision Y, a costly signal, and ignores the statements. Second, conditional on the decision taken, a member uses a statement strategy that is the same for all pairs of signals. VS draw a plausible, second conclusion that is, however, not dictated by game-theoretic logic: members will show a united front and speak with one voice to the market. Game theory dictates then that the market should be able to assess a member in the out-of-equilibrium event that the committee were not to show a united front. It is consistent with the model to assume that disagreement leads to a drop in assessment. In sum, the model leads to the following predictions. If members care about their assessments, $\lambda > 0$,

1. Evaluators base assessments exclusively on Y and β and ignore ω. Assessments satisfy $\hat{\pi}(Y = 1; \beta) > \hat{\pi}(Y = 0; \beta)$.

2. (i) A committee member uses a statement strategy that is the same for all pairs of signals. (ii) In fact, members form a united front.

3. (i) If both committee members receive the same signal, they take the decision that corresponds with these signals; (ii) If they receive conflicting signals, the committee chooses $Y = 1$ with probability $\beta \in (0, 1/2)$ that satisfies Eq (1).

For what follows, it will be instructive to compare these main predictions with those for the case that strategic interactions between the committee and an evaluator are absent, *i.e.*, $\lambda = 0$.

1′. (i) An evaluator's assessment after $Y = 1$ is independent of ω. It may or may not depend on ω after $Y = 0$, depending on the statement strategy. As-

sessments satisfy $\hat{\pi}(Y=1) > \mathbb{E}[\hat{\pi}(Y=0)]$;[8] (ii) The difference $\hat{\pi}(Y=1) - \mathbb{E}[\hat{\pi}(Y=0)]$ is larger than the corresponding difference $\hat{\pi}(Y=1;\beta) - \hat{\pi}(Y=0;\beta)$ for $\lambda > 0$.

2′. The statement strategy is undefined as it is payoff-irrelevant.

3′. (i) If both committee members receive the same signal, they take the decision that corresponds with these signals; (ii) if they receive conflicting signals, they choose $Y = 0$.

These predictions hold on the equilibrium path. A more basic prediction – an assumption, really – is that committee members and the evaluator best reply to each other's behavior, also off the equilibrium path. Finally, although it is not central to either theory or experiment, theory predicts that members share their private information as conflicts among members are, by construction, absent.

3 The experiment

We begin by describing the *A-Stm* treatment. In this treatment, a committee member cares about his assessment (A) and must send a statement (Stm) to the evaluators. At the start of each session, and before assigning roles to subjects, we handed out written instructions that covered both roles and went through those instructions verbally. Next, the computer randomly assigned half of the subjects the role of committee member and the other half the role of evaluator.[9] Our matching schedule needs to balance two goals. The first goal is the avoidance of uncontrolled dynamic incentives that interfere with the controlled incentives. The second goal is the creation of a common frame of reference in which committee members can identify a relationship between their observable actions and the resulting assessments, and evaluators can understand the meaning of cheap-talk statements. The first goal favors a perfect stranger matching, the second goal favors stable matches that are permanent over time. We therefore chose an intermediate form of rematching of committee members and evaluators. In particular, the computer randomly formed two-member committees and assigned four evaluators to the two members of two randomly chosen committees. The assigned roles, the committees and the matching between committees and evaluators remained the same throughout the experiment.

[8] The assessment after $Y = 1$ is independent of ω as $Y = 1$ reveals that $s_1 = s_2 = s^g$. The statement strategy used in case of $Y = 0$ may or may not reveal the signal pair. We thus write $\mathbb{E}[\hat{\pi}(Y=0)]$ where the uncertainty is about ω.

[9] In the experiment, committees were called groups and committee members decision makers.

Depending on the number of subjects during a session, members and evaluators were matched using a 2×2–scheme or a $2 \times 2 \times 2$–scheme, see Figure 1. This allows for a sufficiently stable relationship between members' actions and assessments to determine a strategy. Fixing the matching within a committee has the additional benefit of realism. Besides, it reduces the time members spend greeting each other and developing a common reference frame of the experiment. This speeds up the experiment. To prevent identification of subjects and the risk of uncontrolled dynamic incentives, in every round the software randomly determined the actual evaluators that were behind the labels 'evaluator 1' - 'evaluator 4' on each committee member's screen. Similarly, the actual committees behind the labels 'committee 1' and 'committee 2' and the actual members behind 'member 1' and 'member 2' for each committee were randomly determined in every round.

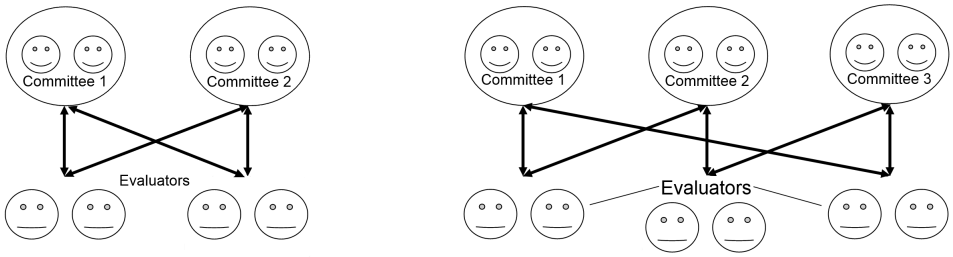

Figure 1: Matching schedules used.

A crucial feature of models of reputation-concerns, like that of VS, is that a member receives a signal about the state of nature and the quality of that signal depends on his (unknown) ability. We used Figure 2 in the instructions – written and on-screen – to explain this relationship in our committee setting. This one picture summarizes all random draws done in a period. At the top, a jar containing two balls, a blue and a yellow one, represents the prior uncertainty about the state of nature. The blue ball represents the bad state, the yellow ball the good state. Next, the computer draws a ball. Either ball has an equal chance of being drawn. As the state is not shown to any subject, the color of the drawn ball is grey. Below the grey ball, there are two columns of boxes, one column for each member in a given committee. Each column consists of three boxes. For each member, two out of three boxes are labeled H to indicate they contain high quality information. Each of these boxes is filled with two balls of the same color as the ball drawn from the jar. One of the three boxes is labeled L to indicate it contains low quality information. This box contains a blue and a yellow ball. Next, the computer randomly selects one of the three boxes and out of this selected box one ball is drawn at random, all with equal probabilities. Thus, the prior likelihood that a committee member

received high quality information, or, equivalently, is of high ability, equals $\pi = 2/3$. In each round in the experiment and for each committee, the computer determined the state of nature, the quality of the information each member receives, and the actual signal that each member received. The computer did not reveal the color of the ball drawn from the jar nor the letter on the box.

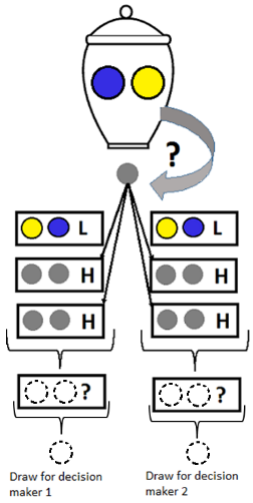

Figure 2: Graphical depiction of relationship between private signals on the one hand and state of nature and ability levels on the other.

After receiving their private signals, members could use a chat window for free-form communication within the committee. Communication was private, *i.e.* remained unobserved by any other participant in the experiment. We chose free-form communication to add realism and to obtain a database that can be studied for, e.g., reasons to behave in a particular way.

Next, a member voted in favor of *Yellow* ($Y = 1$ or implementation) or *Blue* ($Y = 0$ or rejection) as the decision. The committee's decision was $Y = 0$ unless both voted for $Y = 1$. On the next screen, members observed both votes cast and the resulting decision. On that screen, a member was prompted to state his degree of confidence in the decision taken by the group. Possible statements were 'Very Doubtful,' 'Doubtful,' 'Neutral,' 'Confident' and 'Very Confident.' As we want to analyze any effect these cheap-talk statements have on evaluators' assessments, we chose a form of statement that could readily be used in later econometric analysis, rather than free-form communication. This screen also had a chat window for free-form, private communication within the committee. We refrained from prompting members to use the chat window as one of the goals of the experiment was to find out whether different treatments led to different behaviors, including the use of the chat window to discuss, say, assessments and coordinate statements to form a united

front.

Next, the committee's decision and the statements made by each member were presented to four evaluators. Each evaluator was asked to assess, on a scale from 0 to 100%, the chance that a given member had received high quality information in that round.[10] Once each evaluator had assessed the four members, each member observed the state of nature, his committee's decision, the resulting project payoff and the assessments for that round. The project payoff of $Y = 0$ equaled zero, independent of the state, while the payoff of $Y = 1$ equaled 110 if state and decision matched (yellow ball or $\mu = h$) and -120 if they did not match (blue ball or $\mu = -h$). In terms of the parameters of VS, $p = -5$ and $h = 115$. For these values, $\beta = 0.33$. The assessment payoff of a member in a round equaled the average of the assessments obtained. We incentivized evaluators to report their true assessments by rewarding them using a stochastic scoring rule, as in Hossain and Okui (2013) and Schlag and van der Weele (2013).[11].

On the results screen, an evaluator observed her payoff per committee member and was reminded of the decision of both committees, members' statements and her own assessments. The identities of the committee and of its members behind the labels on this screen were the same as on the screen where she provided her assessments. Across rounds, however, the identities were randomly determined.

We designed two more treatments. The *A-NoStm* treatment proceeded as the *A-Stm* treatment with one exception: after members had taken a decision, they did not send statements to evaluators. Thus, evaluators only observed the decision of the committee before they were asked to assess members.

The *NoA-Stm* treatment captures the situation without strategic interaction between committee members and evaluators. To avoid any effect stemming from the presence of evaluators on committee members, we first ran sessions for committee members. Their instructions did not refer to evaluators, and their payoffs equaled the project payoffs. As in the *A-Stm* treatment, once they had taken a decision and before they learned their project payoff, members were prompted to state their degree of confidence in their decision. We used the data so obtained to run sessions for evaluators a few days later. As in the other treatments, their instructions included the instructions we had given to committee members and explained that it was their role to assess these members. Next, we provided them with the actual decisions and statements and they were prompted to submit their assessments as in the *A-Stm* treatment. Their incentives were as in the other two treatments.

[10] In the instructions the expression "received high quality information" was always accompanied by "received a ball from a box labeled H."

[11] See section A.1 for details.

Before the actual experiment began, subjects had to answer questions about the payoffs and probabilities to check their understanding of the set-up. After all subjects answered all questions correctly, the actual experiment began. The experiments took place in the econlab at Erasmus University Rotterdam. All subjects were invited via the econlab subject pool using ORSEE, see Greiner (2004). The experiments were programmed in php/my-sql and ran on an external server. In total, subjects completed 17 rounds in the *A-Stm* and *A-NoStm* treatments and 32 rounds in the *NoA-Stm* treatment. The larger number of rounds in the latter treatment helped to equalize the duration of the three treatments.[12] In all sessions, the first two rounds were practice rounds that could not be selected for payments. Subjects were instructed to use these rounds to get acquainted with the computer environment and the task. In what follows, the first two rounds are dropped from the data before analysis, unless explicitly stated otherwise. At the end of the experiment, the computer randomly selected four rounds for payment in the *A-Stm* and *A-NoStm* treatments. In the *NoA-Stm* treatment, we selected four rounds for payment for the evaluators, but ten rounds for committee members. The higher number of rounds was needed to compensate members for the absence of assessments in their payoffs. We chose ten rounds as the expected total payoff of predicted committee behavior for this number of rounds is about equal to the expected total payoff (including assessments) of predicted behavior in the other two treatments over four rounds. Earnings for these rounds were added to the show-up fee, €5. After the experiment, subjects filled out a questionnaire about some background characteristics, before getting paid in cash and leaving the lab. Sessions lasted about 1 hour and 45 minutes, including instructions and payment. On average, subjects earned €21.26, approximately $28 at the time of the experiments.

Table 1 shows, for each treatment, the number of subjects that participated and the resulting numbers of decisions, statements and assessments. In total 224 subjects participated in our experiment, 88 in the *A-Stm* treatment, 80 in the *NoA-Stm* treatment, and 56 in the *A-NoStm* treatment.

Table 2 presents some characteristics of the subjects. About half of the subjects is male and they are about 21 years of age. A majority studies economics or business and they have studied for 2.6–3 years. We also asked subjects to respond to the general risk question used in Dohmen et al. (2011). The answers range from 1, not at all willing to take risk, to 11, very willing to take risk. On average, they score 5.4. We test for the similarity of the distributions of the characteristics through χ^2-tests.

[12] In section A.5, we show that results are qualitatively the same in the first and second half of the experiment.

Table 1: Number of subjects, decisions, statements and assessments

Treatment	CM	EV	Decisions	Statements	Assessments
A-NoStm	28	28	210	-	1,680
A-Stm	44	44	330	660	2,640
NoA-Stm	38	42	570	1,140	5,040
Total	110	114	1,110	1,800	9,360

Notes: Number of observations of the most important variables. Each round, in all treatments two committee members (CMs) take a decision together and, in the *A-Stm* and *NoA-Stm* treatments, each CM makes a statement. Every evaluator (EV) assess 4 CMs per round. There are 15 rounds in the *A-Stm* and *A-NoStm* treatments and 30 rounds in the *NoA-Stm* treatment.

Table 2: Subject characteristics

	Treatment			
	A-NoStm	A-Stm	NoA-Stm	χ^2
Male	0.50	0.60	0.58	pr=0.475
Year	2.60	3.10	3.01	pr=0.097
Econ background	0.68	0.82	0.86	pr=0.026
Age	20.96	21.32	21.43	pr=0.276
Risk tolerance	5.39	5.41	5.56	pr=0.425

Notes: Distribution of subject characteristics and χ^2-test of equality of the distribution per variable. *Male* is a dummy set to 1 for male participants. *Year* is the year of the study the subject is in. *Econ background* is a dummy set to one for students in Economics or Business. *Age* is the age of the subject in years. A subject's *Risk tolerance* is measured by the subject's answer to the general risk question used in Dohmen et al. (2011). The answer 1 is not at all willing to take risk and 11 very willing to take risk.

These tests show that there is a relatively small number of economics students in the *A-NoStm* treatment. As we show in appendix A.2, this does not qualitatively affect the results. Since all subjects were recruited in the same way and we always ran several treatments in any given week, we have no explanation for this difference.

Finally, the chat conversations were independently coded by two research assistants according to a common coding scheme.[13] The coding scheme asked them to indicate whether members report their private signals, whether they discuss what decision to take or what statement to send, whether they related assessments to decisions or to statements etc. The two sets of coded conversations were compared and differences resolved by the research assistants.

[13]See section A.8.2 for the coding scheme.

4 Behavior observed in the lab

We begin with a discussion of the evaluators' assessments as they form an important part of the payoffs of committee members in two treatments.[14] We then turn to the chat among committee members to see whether they discuss what relationship, if any, they believe to hold between their observable actions and the assessments they obtain. Next, we analyse committee members' behavior. Section 5 compares the theory with the findings.

4.1 Evaluators

We use OLS regressions to determine the weights that evaluators attach to the decision and, depending on the treatment, the statements that committee members make. Since committee members could choose from five statements, one can control for them in several ways.

Table 3: Statements as observed by evaluators

	A-Stm		NoA-Stm	
Statement	Frequency	Percentage (%)	Frequency	Percentage (%)
VD	4	0.15	32	0.63
D	88	3.33	368	7.30
N	292	11.06	904	17.94
C	416	15.76	2,344	46.51
VC	1,840	69.70	1,392	27.62
Total	2,640	100	5,040	100

Notes: *Frequency* counts the number of evaluator-periods in which an evaluator observes a particular statement. *Percentage* expresses that number as a percentage of the total number of evaluator-periods.

As Table 3 shows, evaluators rarely observe statements that explicitly state doubt, especially if committee members care about assessments. It is thus of little use to control for statements below 'Neutral.' In the regressions we therefore control for the statements 'Very Confident' and 'Confident' using dummies, while the statements 'Neutral,' 'Doubtful' and 'Very Doubtful' are grouped together and used as the low-confidence comparison group. The most extensive model that we estimate

[14] In the *A-Stm* and *A-NoStm* treatments, the average per-period assessment that a committee member receives equals 68.4 and 60.0, respectively. This should be compared with a decision payoff equal to either 0, 110 or −120.

is

$$A_{ijt} = \alpha + \gamma_1 D\left(Y_{it} = 1\right) + \gamma_2 D\left(\text{stm}_{it} = VC\right) + \gamma_3 D\left(\text{stm}_{it} = C\right) + \gamma_4 D\left(\text{Same Statement}_{it}\right) + FE_t + FE_j + \epsilon_{ijt}, \quad (2)$$

where A_{ijt} is the assessment by evaluator j of member i in period t, α a constant, $D\left(Y = 1_{it}\right)$ a dummy that is 1 if the committee of which i is a member chooses $Y = 1$ in period t, $D\left(\text{stm}_{it} = VC\right)$ a dummy that equals 1 if member i in period t uses statement 'Very Confident,' $D\left(\text{stm}_{it} = C\right)$ a dummy that equals 1 if member i in period t uses 'Confident,' $D\left(\text{Same Statement}_{it}\right)$ a dummy that equals 1 if both members in the committee that i is part of chooses the same statement in period t, FE_t are period fixed effects and FE_j evaluator fixed effects, and ϵ_{ijt} a zero-mean disturbance term. Table 4 reports the estimates. In (1), the *A-NoStm* treatment, statement-related dummies are excluded as no statements were made. For comparison, and because theory predicts that assessments are only based on the decision Y, columns (2) and (3) exclude statements from the regressions in the other treatments as well.

Columns (1)–(3) show that in all treatments, evaluators reward the decision $Y = 1$ with a higher assessment than $Y = 0$. The difference is particularly large in the *A-NoStm* treatment, where the decision is the only source of information that evaluators have.

If we control for statements in the *A-Stm* and *NoA-Stm* treatments, the effect of $Y = 1$ becomes smaller, and statistically insignificant in the *A-Stm* treatment. In both these treatments, cheap-talk statements determine to a large extent how evaluators assess committee members. Indeed, the coefficients on 'Very Confident' and 'Confident' are considerably larger than the coefficient of $Y = 1$. Given their frequent use, see Table 3, these large coefficients can best be understood to mean that *deviating* from them implies a substantial drop in assessment. Evaluators seem to believe that if a member expresses doubt about his decision, he is likely to have received a low quality signal.[15] Using the same statement as the other group member implies a somewhat higher assessment.

A comparison of columns (1) and (4) shows that evaluators shift their attention from decisions to statements when the latter are made available to committee mem-

[15]Note that the committee can use the statements to signal at most three signal pairs, namely two negative, two conflicting, or two positive signals. As a result, combining the lowest two statements does not lead to a considerable loss of information. As a consistency check we re-ran the regressions using the statements as a continuous variable, as well as with separate dummies for all but one possible statement (not reported). These regressions confirm that the effect of statements on assessments is monotone.

Table 4: Assessments as a function of observables

VARIABLES	(1) Assessment	(2) Assessment	(3) Assessment	(4) Assessment	(5) Assessment
$Y = 1$	9.602***	4.556*	7.545***	2.222	4.631***
	(1.605)	(2.110)	(1.571)	(1.874)	(1.318)
Very Confident				17.07***	26.97***
				(2.614)	(2.866)
Confident				12.57***	16.61***
				(2.669)	(2.371)
Same Statement				2.466**	1.610*
				(0.760)	(0.870)
Constant	54.04***	64.14***	57.56***	51.98***	43.38***
	(1.463)	(3.686)	(2.025)	(3.329)	(2.077)
Observations	1,680	2,640	5,040	2,640	5,040
R^2	0.518	0.295	0.285	0.407	0.569
Cl-level	match	match	match	match	match
Clusters	6	10	21	10	21
Subject FE	YES	YES	YES	YES	YES
Period FE	YES	YES	YES	YES	YES
Treatment	A-NoStm	A-Stm	NoA-Stm	A-Stm	NoA-Stm

Notes: *Assessment* is the assessment given by an evaluator for a particular member, on the original 100-point scale. $Y = 1$ is a dummy set to 1 if this member's committee chooses $Y = 1$. *Very Confident* is a dummy set to 1 if the member uses the corresponding cheap-talk statement (similarly for *Confident*). *Same Statement* is a dummy set to 1 if this member uses the same cheap-talk statement as his fellow committee member in that period. Fixed effect specification. Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

bers as an instrument to manage their assessments.[16] A comparison of columns (4) and (5) shows that the coefficients on the decision and the statements are smaller and the constant is larger in the *A-Stm* treatment than in the *NoA-Stm* treatment (F-test, $p = 0.0167$).[17] In other words, evaluators make their assessments less responsive to the observed behavior of committee members when the latter are known to care about these assessments. Finally, in section A.3 we show, using regression including interaction terms, that evaluators treat decisions and statements as separate sources of information.

We can thus draw two conclusions. First, a comparison of columns (4) and (5) shows that evaluators are aware that members strategically use their behavior to obtain strong assessments. Second, a comparison of columns (1) and (4) shows that evaluators shift their attention from committees' decisions to their statements when the latter are made available as an instrument to manage assessments.

4.2 Committee members

In two treatments, assessments form an important part of the payoffs of committee members. We begin by studying conversations among members to see whether they discuss assessments and if so, what they believe is driving them. Then we turn to the consequences these beliefs have for decision making and statements used.

4.2.1 Beliefs and belief formation as obtained from the chat

Committee members use the chat boxes to exchange their private signals, to decide what to vote and, in two out of three treatments, what statement to use.

Table 5 presents some key features of the chat.[18] The top part of Table 5 shows almost all members in every round in the *A-NoStm* and *NoA-Stm* treatments chat about the private signals they have received. The percentage is considerably lower in the *A-Stm* treatment. This difference is partly due to the fact that in this treatment there is a comparatively large group of members who, rather than saying what signal they have received, immediately tell what vote they intend to cast. This is clear from a comparison of the percentages in the first and third line of the table. As we show in section A.8, members who chat about their private signals or immediately jump to telling what they intend to vote virtually always truthfully reveal their signals.

[16] A test of difference of coefficient of $Y = 1$ in the two treatments yields $p < 0.01$.

[17] In a test of difference of individual coefficients in the two treatments, only the difference of the *Very Confident* coefficient is significant ($p < 0.01$).

[18] Table A.18 presents a complete overview of all coded variables.

On average, members write less often about what to vote than about their private signals. They quickly agree that if they receive the same signal, they vote in line with those signals. In case of conflicting signals, discussions about what to vote remain common. By and large, writing about statements is absent from the *NoA-Stm* treatment. This was to be expected given their payoff-irrelevance. In the *A-Stm* treatment, most committee members write about statements, but the frequency goes down over time.

The lower part of the table shows among others what members believe to be driving the assessments they receive.[19] In particular, it shows that the inclusion of cheap-talk statements as an instrument to influence assessments leads to a marked shift in the discussions among members as to what shapes assessments. Members in 11 out of the 22 committees in the *A-Stm* treatment relate at some point during the experiment the assessments they receive to the statements they make. In this treatment, only 2 committees ever relate the assessments they receive to the decisions. As a result, members discuss much more often what statements to choose in a bid to affect their assessments than what decisions to take. The emphasis on statements is justified by the way evaluators assess members, see section 4.1. In the *A-NoStm* treatment, 5 out of the 14 committees discuss the relation between assessments and decisions, typically pointing out – correctly – that evaluators reward $Y = 1$ more than $Y = 0$.

The chat of the committees that discuss the relationship between their statements and assessments sheds light on how committees come to form these beliefs. Consider the *A-Stm* treatment. Typically, after the first rounds have past, a member shares with his partner his finding that statements of confidence lead to higher assessments than statements of doubt and suggests to fool evaluators by choosing '(Very) Confident' even though they received conflicting signals. The other member agrees and right at the beginning of the next round they share their joy over their success. Galvanized by this experience, they use statements expressing confidence much more frequently or even all the time. Rarely do committee members take a more 'cerebral' approach to understanding the relationship between their actions and resulting assessments.[20]

Articulating the belief that statements affect assessments changes members' behavior. This is illustrated in Figure 3. To make this figure, we score every statement

[19]The statistics in the top part of the table are based on the incentivized rounds, whereas the numbers in the lower part are also based on the first two practice rounds. We include the practice rounds as part of the understanding of the game as reported in the lower panel may take place in these rounds.

[20]See excerpts 1 and 2 in section A.8.1 for two committees that use past experience to come to the belief that statements shape assessments. Excerpt 3 illustrates the cerebral approach.

Table 5: Chat – summary statistics

	A-NoStm	A-Stm	NoA-Stm
Percentage of member-rounds with messages about:			
- private signal received	99.5	77.0	99.7
- vote in that round	76.7	52.4	52.6
- signal received or vote in that round	100	83.6	99.8
- statement in that round	–	32.1	4.4
Number of committees in treatment	14	22	19
Number of committees discussing:			
- link between statements and assessments	–	11	–
- link between decisions and assessments	5	2	–
- risk taking	12	16	19
- zero-payoff dilemma	3	3	9
- united front	–	2	–

Notes: Summary of topics of discussion in the chat. Each committees conversation in the two chatboxes in a round were treated as a single observation.

on a 5-point scale, from 1 for 'Very Doubtful' to 5 for 'Very Confident.' The figure presents the average statement scores for two types of members in specific rounds. The green dots, labeled 'Not discussed,' represent the average statement score of members in committees that have not discussed the relationship between statements and assessments before or during the round that his committee received conflicting signals for the nth time. Note that n varies along the horizontal axis. The red diamonds, labeled 'Discussed,' represent the average statement score of members in the remaining committees, in which the relationship has been discussed before or during the round in which it receives conflicting signals for the nth occasion. The figure shows that relating assessments to statements leads members to raise their levels of stated confidence.

Whether to take risk, *i.e.*, to choose $Y = 1$ in case of conflicting signals, is a common topic of conversation in all treatments. In some committees, conflicting signals also leads to a different, but related discussion: by choosing $Y = 0$, members exclude the chance of receiving a positive project payoff and receive 0 for sure. This 'zero-payoff dilemma' is especially felt if, as is the case in the *NoA-Stm* treatment, committee members' project payoffs are their only payoffs. In the other two treatments, this dilemma is little discussed, probably because the stark contrast between zero points for sure and the possibility of earning a positive number of points is absent thanks to the presence of (positive) assessments irrespective of the decision. These discussions also have behavioral consequences. In the *NoA-Stm* treatment, committees that haven't discussed this dilemma choose $Y = 1$ in 19.9 % of all pe-

riods with conflicting signals. Once committees have discussed the dilemma, this percentage jumps to 61.3 % ($p < 0.001$ in a two-sided t-test with unequal variances.)

In the theory of VS, a united front is the result of a conscious choice to act in tandem, not a coincidentally appearing equality of statements used vis-à-vis evaluators. This is clear from VS's use of the phrase by Frederick H. Schultz, a former Governor and Vice-Chairman of the FOMC, "[w]e should argue in the Board meetings but close ranks in public" (VS, p. 339) to illustrate a united front. A proper test of the theory can therefore not simply count how often members choose the same statement. Instead, we count the number of times a committee member bring up the importance of a united front in a Schultz-like manner.[21] Only two committee do, see excerpts 4 and 5 in section A.8.1.

The next two subsections analyze differences in behavior across treatments. In section 4.2.4, we discuss behavioral heterogeneity within treatments.

4.2.2 Consequences for decision making.

A belief that assessments are related to statements rather than to decisions, as in the *A-Stm* treatment, should lead to a lower frequency of distorted decisions ($Y = 1$ in case of conflicting signals) than in the *A-NoStm* treatment. Figure 4 shows, for each treatment and for a given number of positive signals s^g that a committee received, the fraction of $Y = 1$-decisions. In case of conflicting signals, $Y = 1$, is chosen 25% of the time in the *A-Stm* treatment, a percentage that is indeed significantly lower than the 37% in the *A-NoStm* treatment (two-sided test of proportions, $p = 0.025$). Also note that the rewards in terms of a stronger assessment are larger in the *A-NoStm* treatment than in the *A-Stm*, see the coefficients on $Y = 1$ in Table 4.

The chat shows that the 'zero-payoff dilemma' is especially felt in the *NoA-Stm* treatment. It has an important consequence for committee behavior in that treatment compared with their behavior in the *A-Stm* treatment. In the latter treatment, this dilemma is absent and assessments are associated with statements, hardly with decisions. It would thus be consistent with the discussions in the chat to observe that more committees distort the decision on Y in the *NoA-Stm* treatment than in the *A-Stm* treatment. Figure 4 shows that this is indeed the case. Conditional on conflicting signals, committees whose members only care about decision payoffs choose $Y = 1$ 34% of the time, which is significantly more than when members also care about their assessments (two-sided test of proportions, $p = 0.052$).[22]

In sum, the beliefs, as obtained from the chat, about the determinants of as-

[21] We prefer this test over one based on a sentence – common in the *A-Stm* treatment – like "shall we choose confident?" as it would lack an articulation of the importance of using the same

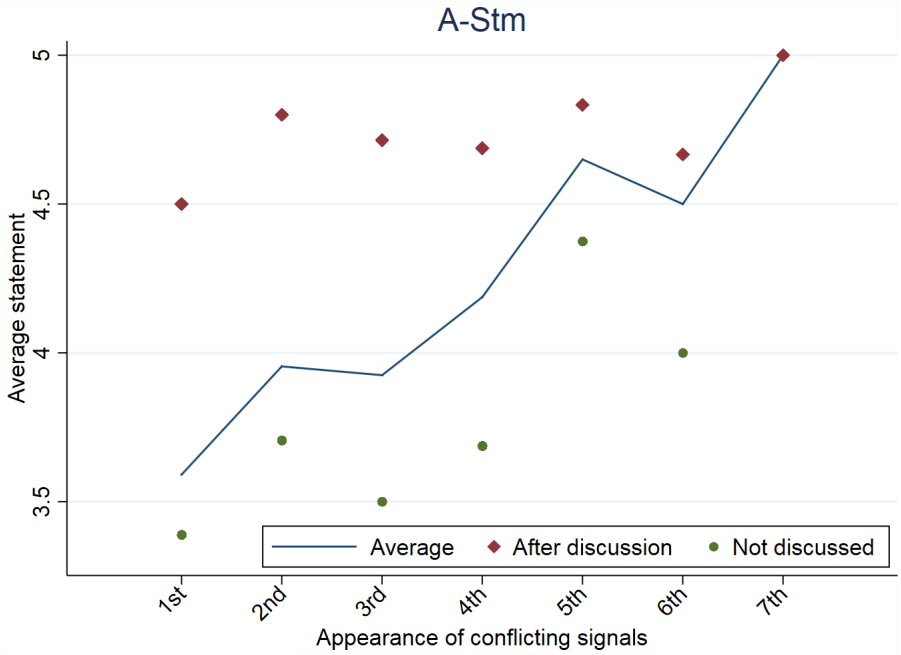

Figure 3: Discussing the link between statements and assessments increases the stated level of confidence

Notes: The average statements made by members in specific rounds. The green dots represent the average statement score of members in committees that have not (yet) discussed the relationship between statements and assessments before or during the round that his committee received conflicting signals for the nth time. The red diamonds represent the average statement score of members in the remaining committees, in which the relationship has been discussed before or during the round in which it receives conflicting signals for the nth occasion. The blue line is the overall average statement made in the nth round of conflicting signals.

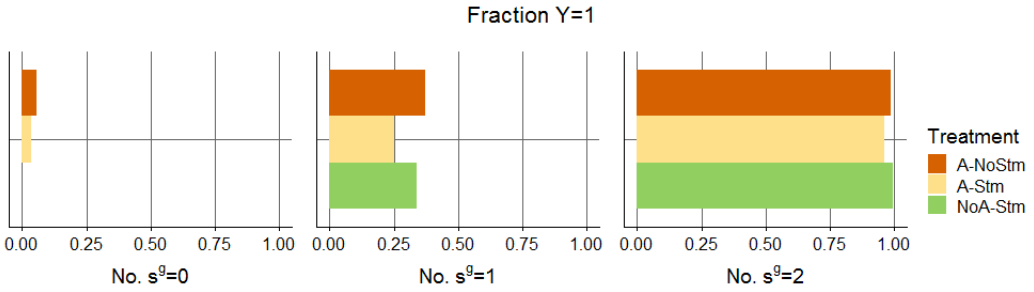

Figure 4: Relationship between number of positive signals and committee decision

Notes: Each panel shows, for each treatment and for a given number of positive signals, the fraction of committee-rounds with $Y = 1$.

sessments have consequences for the decisions taken. The presence of cheap-talk

statement.

[22]Incidentally, the figure also shows that, if both members receive the same signal, with few exceptions, committees take the decision that matches their signals.

statements changes members' beliefs about what drives assessments and leads to a reduction in the frequency with which the decision is distorted. Furthermore, the absence of assessments as part of the incentives leads to an increase in the share of committees discussing the zero-payoff dilemma and a concomitant increase in the frequency of distorted decisions.

4.2.3 Consequences for statements

In the *A-Stm* treatment, members relate the assessments they obtain to the statements they make, hardly to the decision they take. As a consequence, they use a markedly different statement strategy than members who only care about decision payoffs. Both the language used and the meaning implied change.

Figure 5 shows the language used. It reports the relative frequency with which committee members use the various statements conditional upon the signals they have received and the decisions they have taken. In the absence of a concern with assessments, frequency distributions are roughly bell-shaped. When they have received the same signals, members' modal statement is 'Confident.' They use this statement 62% of the time. The next most common statement, 'Very Confident,' is used less than half of that. Statements expressing less or no confidence are hardly used. After conflicting signals, and depending on the decision taken, 'Neutral' or 'Doubtful' becomes the modal statements, used close to 45% of the time.

A concern with assessments shifts the distribution of statements to the right. 'Very confident', the most extreme statement, becomes the modal statement that members use, irrespective of the pair of signals they have received. They use this statement around 80% of the time if their signals agreed, and more than 40% of the time if they have received conflicting signals. Statements expressing doubt become extremely rare.

The interpretation of a statement – *i.e.*, what a statement says about the ability of a member – also changes due to a concern with assessments. Table 6 reports for each statement the fraction of committee members who had conflicting signals amongst members using this statement. As conflicting signals are a sure sign that at least one committee member received low quality information, the higher this fraction is, the higher is the likelihood that a member is of low ability. The table shows, for each decision and in both treatments with statements, a monotone relationship between this fraction and the degree to which a statement expresses confidence in the decision taken. It also shows that a concern with assessments decreases the variation in this fraction across the statements that members use, especially for $Y = 1$. In other words, in the presence of a concern with assessments, statements become

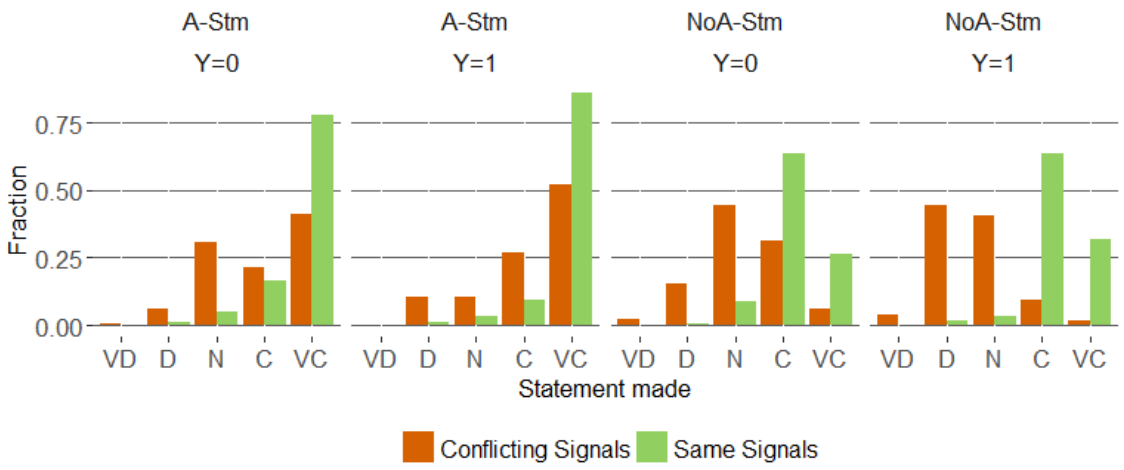

Figure 5: Observed statement strategies

Notes: The figure shows, for each combination of treatment, decision and signal pair, the fraction of member-rounds with a certain statement. As committees rarely vote for $Y = 1$ when they receive (s^b, s^b) signals, and vice versa for $Y = 0$ with (s^g, s^g), we dropped those observations for clarity of presentation.

more similar and reveal less about members' ability. We return to this observation in section 6.

Table 6: Likelihood that committee members received conflicting signals conditional on a statement

	$Y = 1$		$Y = 0$	
Statement	A-Stm	NoA-Stm	A-Stm	NoA-Stm
Neutral/Doubtful/Very Doubtful	0.4	0.84	0.76	0.79
Confident	0.35	0.04	0.46	0.23
Very Confident	0.10	0.02	0.26	0.13

Notes: Number of committee-rounds with conflicting signals as a fraction of total number of committee-rounds for which a member uses the indicated statement, per treatment and decision.

We noticed that members form their beliefs over time and that beliefs shape behavior. This has consequences for the distribution of statements in the *A-Stm* treatment. In the first 5 (of 15) rounds, committee members use 'Confident' or 'Very Confident' some 58% of the times they receive conflicting signals and 92% of the times they have the same signals. In other words, many members start out by using statements that depend on the signals they have received. As members realize that higher confidence statements lead to higher assessments, in the last 5 rounds, the frequency with which members report 'Confident' or 'Very Confident' has gone

up to 78% in case of conflicting signals, with 'Very Confident' causing the bulk of the increase. Conditional on having received the same signals, the usage frequency of these two statements remains constant, at 93%. The only change that is happening for this signal pair is that 'Very Confident' becomes even more dominant than it already was. These changes only happen in the *A-Stm* treatment, not in the *NoA-Stm* treatment, which shows they are caused by a concern with the assessments.[23]

4.2.4 Behavioral heterogeneity within treatments

A focus on average behavior per treatment may hide differences between committee members within a treatment. In the experiment, members' behavior varies in two important respects. They differ in the frequency with which they vote v^1 if the committee they are part of has received conflicting signals. We call this frequency a member's *inclination to distort*.[24] And they show variation in the extent to which their statements after conflicting signals differ from those they choose after concurring signals. To measure this variation, we score every statement on the 5-point scale introduced above. We calculate for each member two average statement scores, one averaging over all rounds in which he received the same signal as his fellow member and one over all rounds with conflicting signals. We call the difference between these two scores a member's cheap-talk *transparency*. The absolute value of this difference runs from 0 to 5, with 0 meaning that, on average, a member uses the same statement in both types of rounds.

Figure 6 shows the distribution of members' inclination to distort in case of conflicting signals. These distributions are quite similar in the *A-Stm* and the *A-NoStm* treatments. In either treatment, the modal inclination is zero. In other words, the modal response of members is to vote $v = v^0$ any time they receive conflicting signals. Moreover, there is considerable variation in members' inclinations. In the *NoA-Stm* treatment, the modal inclination is close to 1/2 and the presence of members who don't distort at all is significantly smaller than in the other two treatments.

Figure 7 shows the distribution of members' transparency. The modal transparency score in the *A-Stm* treatment is zero: more than 30% of committee members make it impossible for evaluators to glean information about the signal pair they have received from the statements they use. The distribution in the *NoA-Stm*

[23] In the *NoA-Stm* treatment, neither percentage changes significantly. The confident statements are used 30% of the time conditional on conflicting signals in both the first and the last five periods, while these percentages equal 90% in the first period and 94% in the last, conditional on members having received the same signals.

[24] As all committees received conflicting signals in at least one round, this frequency can be determined for every member.

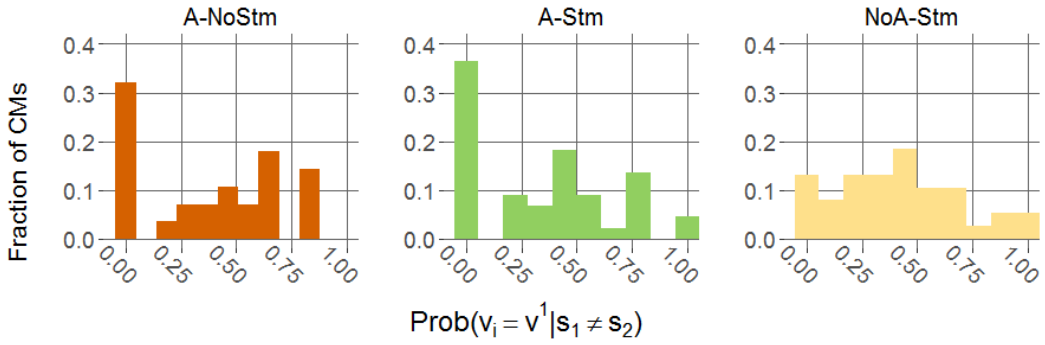

Figure 6: Members' inclination to distort

Notes: A member's inclination to distort is defined as the fraction of periods with conflicting signals in which a member votes $v = v^1$.

treatment is quite different. In fact, the modal difference is 1 full point and many members are even more transparent.

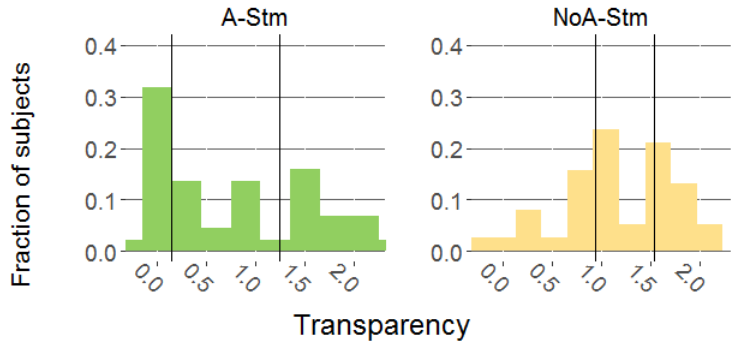

Figure 7: Members' cheap-talk transparency

Notes: A member's cheap-talk transparency is defined as the difference in the average statement between rounds in which he received the same signal as his fellow member and rounds in which they received conflicting signals. To determine the average statement, we score every statement on a 5-point scale, from 1 for 'Very Doubtful' to 5 for 'Very Confident.' The 33th and 67th percentile of the distribution are marked with vertical lines.

To see whether there is a relationship between a member's inclination to distort and his transparency, we divide members in terciles on the basis of their transparency and then determine for each tercile the number of members with a zero inclination to distort and those with a positive inclination to distort. The result is shown in Table 7. The vertical lines indicate the cut-offs between terciles. In the *A-Stm* treatment there seems to be a division between those committee members that act strategically and have both a low level of transparency and distort their decisions, and those that do not act strategically on either dimension. In the *NoA-Stm* treatment, on the other hand, these tendencies are absent; instead, those who are inclined to distort

the decision are equally present in all transparency terciles.

Table 7: Cheap-talk transparency and distorted decisions

	Cheap-talk transparency tercile					
	A-Stm			NoA-Stm		
Inclination to distort	bottom	middle	top	bottom	middle	top
no	4	8	10	4	3	3
yes	11	6	5	9	9	10
Total	15	14	15	13	12	13

Notes: Joint distribution of members' cheap-talk transparency and inclination to distort the decision. A member's transparency equals the difference in the average statement between rounds in which he received the same signal as his fellow member and rounds in which they received conflicting signals. A member's inclination to distort is defined as the fraction of periods with conflicting signals in which a member votes $v = v^1$. If this fraction equals zero, then inclination to distort equals no; if fraction is positive, then inclination to distort equals yes.

In sum, there is heterogeneity in member behavior. A concern with assessments creates a relationship between the two dimensions of heterogeneity. There are types that are both unlikely to distort their choices and unlikely to exaggerate their cheap-talk statements although they are being assessed, and strategic types that do both because they are assessed.

5 Behavior in the lab versus behavior in theory

Theory assumes that an evaluator best-replies to the decisions (and statements) that she observes. Vice versa, committee members are assumed to best-reply to the assessments of evaluators. In equilibrium, members strategically interfere with the inferences that evaluators draw from the information they receive and trade-off gains in assessments against expected losses in decision payoffs. This gives rise to the predictions summarized on page 9. How well do these predictions correspond to observed behavior in the lab?

A key prediction is that evaluators give higher assessments to committee members, reputation-concerned or not, after $Y = 1$ than after $Y = 0$ (predictions 1 and $1'$, the inequality signs). This prediction describes the average actual assessments well, see columns (1)–(3) in Table 4.

Prediction 2 is that in the *A-Stm* treatment, statements contain no relevant information concerning members' ability. Besides, members consciously form a united front. We find little evidence that committees in the *A-Stm* treatment consciously

formed a united front. We do find that the modal degree of cheap-talk transparency is zero. Moreover, over time, a growing number of members reveals less about their private signals through their statements. A comparison of the average difference in statements between rounds with conflicting and same signals, shows that transparency is much higher in the *NoA-Stm* treatment than in the *A-Stm* treatment (difference of 0.41 on a 5-point-scale, $p < 0.001$ in a t-test). This shows that reputation concerns make statements considerably more uniform across signal pairs. Nevertheless, many members do reveal information about their ability levels through their statements.

Given that the statements of the majority of members do contain information, the relevant theoretical prediction for evaluators is that they best-reply to this information and integrate it in their assessments, rather than ignore cheap-talk statements (prediction 1, the independence of ω). We study whether this is the case in section 5.1.

Prediction 1′, part (ii), says that reputation concerns make evaluators' assessments react less to 'positive' information produced by committees. As a simple test for this prediction, we compare the determinants of evaluations in the *A-Stm* and *NoA-Stm* treatments. Columns (5) and (6) in Table 4 show that, indeed, assessments go up less after statements of confidence and the decision $Y = 1$ when evaluators know that committee members care about receiving high assessments. A better test is, again, to see whether evaluators best-reply to observed committee behavior, the topic of section 5.1.

Theory correctly predicts that if both members of a committee receive the same signal, the committee follows those signals (predictions 2 and 2′, part (i)). The predicted pattern of decisions in case of conflicting signals is not found in the data (predictions 3 and 3′, part (ii)). Indeed, compared with the *NoA-Stm* treatment, adding reputation concerns reduces, rather than increases, the incidence of $Y = 1$. Members' conversations suggest that this inclination to vote for $Y = 1$ in the *NoA-Stm* treatment is due to the zero decision payoff in case of $Y = 0$ looming large. A second reason may be that the observed gain in assessment from choosing $Y = 1$ rather than $Y = 0$ is larger in the *A-NoStm* treatment than in the *A-Stm* treatment. This should increase the number of votes for $Y = 1$ in case of mixed signals in the *A-NoStm* treatment.

The next section uses an orthogonality test that is based on Keane and Runkle (1990, 1998) to shed light on any systematic mistakes made by evaluators in transforming the information they observe into the assessments they come up with. In section 5.2, we study which reputation-concerned committee members performed

best given the observed assessments of evaluators.

5.1 Do evaluators react rationally to observable behavior of committee members?

Prima facie evidence suggests that assessments are broadly consistent with the behavior of committee members. Consider the following three patterns in the assessments, see section 4.1. Assessments are higher if committees choose $Y = 1$ rather than $Y = 0$; in the *A-Stm* and *NoA-Stm* treatments, assessments go down if members don't state to be (very) confident in the decision; and compared with the *NoA-Stm* treatment, assessments react less to observed committee behavior if members care about their assessments. These three patterns seem to be reasonable replies given the behavior of members: committee members with conflicting signals are less likely to choose $Y = 1$; make lower confidence statements; and these effects are more pronounced in the *NoA-Stm* treatment.

Following Keane and Runkle (1990), we break the rational use of information into two components: the assessments have to be unbiased and efficient estimators of ability. Evaluators are said to provide an unbiased estimate of ability if the (unconditional) average of the assessments matches the (unconditional) average ability. Evaluators are said to use information efficiently if all information about ability that evaluators can glean from observed committee behavior is captured by their assessments.

We define a variable h_{it} to have a value of 100 if committee member i in round t is of high ability (received high quality information) and 0 if he receives low quality information in period t. This variable captures ability, and is defined on the same percentage-points scale as the assessments of the evaluators. The test in Keane and Runkle (1990, 1998) involves two regressions, one for unbiasedness, the other for efficiency, each of the form:

$$h_{it} = \alpha_0 + \alpha_1 A_{ijt} + \alpha_2 X_{ijt} + \epsilon_{ijt}, \tag{3}$$

where A_{ijt} is the assessment given to committee member i by evaluator j in period t, and X is either a matrix of explanatory variables that are observable to evaluator j, or empty. Unbiasedness then translates to the restrictions $\alpha_0 = 0$ and $\alpha_1 = 1$ in a regression without X_{ijt}, while efficiency requires that $\alpha_0 = 0$, $\alpha_1 = 1$ and all $\alpha_2 = 0$ in a regression including X_{ijt}. This orthogonality test requires OLS-type regressions where the prediction can be interpreted as a conditional expected value. However, our dependent variable is binary, not continuous. Moreover, running

this test with OLS regressions requires that the error terms be independently and normally distributed. Neither is the case in our setting. To circumvent these issues, we adapt Keane and Runkle's approach to our setting.

To test for unbiasedness, we use the two-sided t-test in Table 8. The t-tests show a large difference in the treatment without statements, *A-NoStm*. The average assessment in this treatment is too low by about 4 percentage points. In the treatments with cheap-talk statements, differences are considerably smaller and even insignificant in case of *A-Stm*. We can therefore conclude that there are no biases in the *A-Stm* treatment, some indications of biased assessments in the *NoA-Stm* and clear issues in the *A-NoStm* treatment. In the latter two treatments, average evaluations appear to be too low. The difference of 4 percentage points in the *A-NoStm* treatment is roughly 6 % of the average assessment – significant, but not extremely large from an economic point of view.

Table 8: Assessments and true ability per treatment

	Assessments		Ability		Two-sided t-test		
Treatment	Mean	Std. Dev.	Mean	Std. Dev.	diff.	Pr(diff.)	Freq.
A-NoStm	60.04	15.41	64.29	47.93	-4.24	0.00	1,680
A-Stm	68.37	16.33	67.58	46.82	0.79	0.40	2,640
NoA-Stm	65.29	17.61	66.90	47.06	-1.61	0.02	5,040

Notes: *Assessments* are evaluators' assessments of committee members' ability. *Ability* is the h_{it} variable. It equals 100 or 0 if member i in round t is of high, or low, ability, respectively. We use a two-sided t-test to test for unbiasedness.

Efficiency implies that the change in conditional expected value of assessment should be the same as the change in the conditional expected value of ability for all observable signals of ability. The observable signals of ability should therefore not change the expected value of the *difference* between the assessment and ability. In our experimental setting, we can measure the difference between assessments and ability by directly taking the difference between these observed variables. In terms of Kean and Runkle's test in Eq. (3), this amounts to assuming $\alpha_1 = 1$ and moving the A_{ijt}-term to the other side of the equals sign. After this transformation we can run the adjusted regressions. Note that differences in level that are not related to the observable signals, like the biases in average assessment identified above, are absorbed by the constant in the regression. Define the variable *Mistake*, denoted by Δ_{ijt}, as $\Delta_{ijt} = h_{it} - A_{ijt}$. The orthogonality test thus takes the following form:

$$\Delta_{ijt} = \alpha_0 + \alpha_2 X_{ijt} + \epsilon_{ijt}. \tag{4}$$

After this transformation, there is still a strong correlation between the information that is available about the two members of a given committee in every period, as they have taken the same decision and face the same state of nature. Similarly, every committee member is evaluated by four evaluators in his matching group, creating a common history within matching groups. Like in Keane and Runkle (1990, 1998), we therefore cannot assume that the ϵ_{ijt} are independent within periods or within matching groups. As these authors show, one obtains a consistent estimate of the variance of the coefficients by clustering the standard errors.[25] For our experiment this implies a cross-sectional cluster on the level of the matching group and a temporal cluster on the period. As we have only a limited number of clusters, we bootstrap these clusters using the wild bootstrap procedure of Cameron et al. (2008). The null-hypothesis is, as in the orthogonality test, that all coefficients except the constant equal zero. These regressions are reported in columns (1) to (3) in table 9.

In the *A-NoStm* and *NoA-Stm* treatments, columns (1) and (3), the constant suggests that evaluators, on average, give biased assessments, as we saw before. The constant is only marginally statistically significant in the *A-NoStm* treatment. The orthogonality tests show no significant coefficients on any of the observable variables. This suggests that evaluators tend to make efficient use of available information.

One possible interpretation of the lack of evidence for informational inefficiencies is that our test is not powerful enough. We therefore also perform the orthogonality test, with the same clusters, including a dummy variable that is set to 1 if committee members receive conflicting signals. If two members of the same committee receive conflicting signals, at least one of them is of low ability. This dummy therefore captures considerable information about the ability of committee members, but evaluators don't observe the pair of signals. The results of these regressions are reported in Table A.11 in the Appendix. The coefficient on this dummy is significantly different from zero in all treatments at the 1% level. This shows that the test has the power to detect unused information about ability in the assessments.

In part, the lack of significance on the coefficients of observable variables is because the standard errors are large due to the two-way clustering. We therefore reproduce these regressions without the clusters in columns (4) to (6). These regressions thus heavily overestimate the independent information in every observation.

In columns (4) to (6) the coefficient on $Y = 1$ is never significant. Even without the clustered standard errors there is no sign of systemic mistreatment of the information in the decisions taken by the committees.

[25]For details about this clustering, see also Cameron et al. (2011).

Table 9: Orthogonality tests, with and without two-way clustered standard errors

VARIABLES	(1) Mistake	(2) Mistake	(3) Mistake	(4) Mistake	(5) Mistake	(6) Mistake
$Y = 1$	-1.375	-0.894	1.940	-1.375	-0.894	1.940
	(3.665)	(1.304e+19)	(5.217)	(2.421)	(1.911)	(1.383)
Very Confident		12.58	2.372		12.58***	2.372
		(1.304e+19)	(5.324)		(2.931)	(1.878)
Confident		9.202	8.285		9.202***	8.285***
		(1.304e+19)	(5.125)		(3.390)	(1.698)
Same Statement		-5.003	1.526		-5.003**	1.526
		(1.304e+19)	(2.926)		(2.256)	(1.460)
Constant	4.893*	-7.217	-4.710***	4.893***	-7.217***	-4.710***
	(2.738)	(1.304e+19)	(1.636)	(1.662)	(2.626)	(1.464)
Observations	1,680	2,640	5,040	1,680	2,640	5,040
R^2	0.000	0.007	0.007	0.000	0.007	0.007
Cl-level	Match & Period	Match & Period	Match & Period	NONE	NONE	NONE
Subject FE	No	No	No	No	No	No
Period FE	No	No	No	No	No	No
Treatment	A-NoStm	A-Stm	NoA-Stm	A-NoStm	A-Stm	NoA-Stm

Notes: Columns (1) to (3) report regressions with two-way clustered standard errors; columns (4) to (6) without clustering. *Mistake* is equal to the difference between true ability, h_{it}, and the assessment of this ability, A_{ijt}, both on a 100 point scale. $Y = 1$ is a dummy set to 1 if this members' committee chooses $Y = 1$. *Very Confident* is a dummy set to 1 if this member uses the corresponding cheap-talk statement (similarly for *Confident*). *Same Statement* is a dummy set to 1 if this member uses the same cheap-talk statement as his fellow committee member in that period. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

In columns (5) and (6) the coefficients of *Very Confident* and *Confident* are significant, but have a sign opposite to the constant in both the *A-Stm* and *NoA-Stm* treatment. To interpret this pattern, it is important to recall that statements different from 'Confident' and 'Very Confident' are infrequent, see Table 3. In the *A-Stm* treatment, an F-test on the restriction that the coefficient on *Very Confident* plus the constant equals zero marginally rejects the null-hypothesis ($p = 0.0265$). The same test for the second most likely statement, 'Confident', gives a p-value of 0.4413. This indicates that the average assessment is (close to) correct for the statements expressing confidence.

In the *NoA-Stm* the modal statement is 'Confident.' An F-test on the sum of the coefficient on *Confident* and the constant marginally rejects the null ($p = 0.0156$), while the sum of the coefficient on *Very Confident* and the constant is not significantly different from 0 ($p = 0.1672$). Given the size of these coefficients,

this indicates that as far as deviations exist, they are mostly in the uncommon low confidence statements. The evaluators' mistakes, as far as found, therefore seem to indicate that evaluators have difficulties interpreting uncommon statements.

The upshot is that differences in evaluators' assessments account quite well for the information about committee members' ability level that is contained in their observed behavior. As far as systematic deviations from rational use of information exist, they appear to occur in the *average* level of the assessments. Although these differences in levels could be important for committee members *overall* payoff, their choices are predicted to be determined by *differences* in assessment.

5.2 Which committee members manage their payoffs from assessments and decisions best?

In theory, committee members are not passive bystanders of the process of assessment formation, but are strategically interfering with the inferences that evaluators draw from the information they receive. We have found ample evidence for this prediction. In the *A-Stm* treatment, in which assessments are commonly related to the statements they choose, members statements are markedly different from those used when they are unconcerned with assessments. When we exclude statements as a means to influence assessments, members' attention shifts to decisions and the incidence of distortionary decisions goes up.

At the same time, there is considerable heterogeneity among members. This naturally raises the question which members do best in trading off any gains in assessments against expected losses in decision payoffs. In section 4.2.4, we characterized committee members by their inclination to distort and their degree of transparency. Table 10 shows how decision payoffs and average assessments vary with member's choices on both dimensions (in the *A-Stm* treatment) or only with their inclination to distort (in the *A-NoStm* treatment).

In the *A-Stm* treatment, the least transparent committee members earned the higher assessments, regardless of their inclination to distort. The gain in assessments from distorting the decision in this bottom tercile of transparancy is too small to compensate for the loss in decision earnings. This matches the conclusion based on Table 4: the best strategy for committee members is to always say 'Very Confident' and not distort the decision. Interestingly, during the experiment, various members expressed a concern that if they were to use the same confidence statement time and again, this would raise suspicion with evaluators and possibly harm their average evaluations, see excerpts 6–8 in section A.8.1. The data shows that such a concern

was not only unwarranted, but, to the extent that they kept a statement strategy with a positive transparency score, harmful.

In the *A-NoStm* treatment, the gain from distorting the decision is larger. Given that committees receive conflicting signals in about 1/3 of the periods, an average increase in assessments of 2.4 ($= 60.8 - 58.4$) for all rounds is enough to compensate for the expected loss (conditional on conflicting signals) of 5 in decision earnings in this treatment.

In sum, consistent with the conclusions one can base on Table 4, in the *A-Stm* treatment, the committees that performed best were those that always said 'Very Confident' and refrained from distorting the decision. In the *A-NoStm* treatment, the committees that outperformed the others were those that distorted the decision on Y.

Table 10: Average payoffs received by committee members

	Transparency terciles A-Stm			A-NoStm
	Bottom	Middle	Top	
Inclination to distort, no				
Decision payoff	36.7	39.9	40.3	35.4
Reputation	70.0	66.7	67.4	58.4
Total	106.7	106.6	107.8	93.8
Inclination to distort, yes				
Decision payoff	28.0	28.5	26.5	25.8
Reputation	70.7	69.3	65.1	60.8
Total	98.7	97.8	91.6	86.6

Notes: Breakdown of average per-period payoffs received by committee members, by member's inclination to distort the decision and his degree of cheap-talk transparency in the *A-Stm* treatment, and by his inclination to distort in the *A-NoStm* treatment.

6 Entropy

The orthogonality test presented in section 5.1 examines the extent to which evaluators make rational use of observed committee behavior within a given treatment. Comparisons across treatments are troublesome as these tests don't control for the amount of information that evaluators actually observe in a treatment. As a result, these tests cannot be used to compare the degree to which the information that the computer makes available in each round about members' ability – the pair of members' private signals – is transformed by members' behavior and eventually finds its

way into the evaluations. Such a comparison is, however, relevant to understand how information garbling due to a concern with assessments complicates an evaluator's task, a matter that clearly bears on the possibility of the market for reputation to function properly.

We complement the game-theoretic analysis of behavior with an information-theoretic analysis of the amount of information that is available about a member's ability at various junctures during a round in the experiment.[26] To do so, we measure, on the cardinal entropy scale, the available amount of information in each treatment. Thus, we measure the degree to which the initial uncertainty about a member's ability is reduced by the observed behavior of committees and how much of that reduction finds its way into the assessments that evaluators provide. Because of the cardinal scale, we can make sensible comparisons across treatments and establish, *e.g.*, whether and by how much a concern with assessments reduces the amount of information on which evaluators can base their assessments.

For a random variable X with possible outcomes $x_1, \ldots, x_n$ and associated probabilities $p_1, \ldots, p_n$, the information associated with outcome x_i is defined as $-\log_2 p_i$. It is measured in bits. Thus, the less likely an outcome is, the higher its information. The advantage of using this measure of information is that it places information on a cardinal scale. The entropy of variable X is defined as the expected information of X,

$$H(X) = -\sum_i p_i \log_2 p_i. \tag{5}$$

A binary variable, like a member's ability, has the highest entropy when $p = 1/2$ (it equals one bit). The further away p is from $1/2$, the smaller its entropy becomes, i.e., the less information one expects to receive, or the less uncertainty there is in the variable. For $p = 2/3$, the prior probability that a member is well-informed in the experiment, $H = 0.919$ bit. For $p = 0$ or 1, entropy equals zero as the outcome is known with certainty.

We want to establish how much the initial entropy concerning ability is reduced thanks to the observation of decisions (and possibly statements). Similarly, we want to measure how much information about true ability there is in the assessments. To do so, we need to define the uncertainty – i.e., entropy – that remains about a variable X after observing another variable Z. This is measured by the conditional entropy,

$$H(X|Z) = -\sum_j p_j H(X|z_j). \tag{6}$$

[26]The seminal paper on entropy in information theory is Shannon (1948). A textbook presentation can be found in Luenberger (2006).

The difference $I(X;Z) = H(X) - H(X|Z)$ or

$$I\left(X;Z\right) = \sum_{i,j} p(x_i, z_j) \log_2(\frac{p(x_i, z_j)}{p(x_i)p(z_j)})$$

is the reduction in entropy of X thanks to the observation of Z and is called the mutual information of X given Z. It measures the information about X that is revealed by knowing Z. Within the log, the numerator denotes the probability of observing some joint outcome of X and Z, while the denominator equals the probability of this joint outcome if the variables were independently distributed. This ratio is therefore equal to one, and the reduction in entropy equal to zero, if and only if X and Z are independent. The sum thus takes a weighted average of all probabilities of joint outcomes that occur in a frequency different than would have been expected from independence. In other words, the more the distributions of X and Z depend on each other, the larger is the mutual information and the reduction in entropy of X upon observing Z. One can use this measure to establish how much easier it becomes for a member to determine his ability level once he has observed the signal pair that his committee has obtained, or for an evaluator to predict a member's ability once she has observed a decision and, depending on the treatment, a statement.

Table 11 shows, per treatment, empirical estimates for the initial level of entropy of committee members' ability as drawn by the computer, column (1), and the mutual information of ability given various variables in columns (2)–(6).[27] The binary variable *Confl. Sign.* takes on the value 1 if committee members receive conflicting signals and 0 if they receive the same signals. The binary variable *Y=1* refers to the decision that a group takes, while the *Stm3* variable captures the statements. As before, we bin the lower three statements. *Info_Set* is a variable that combines *Y=1* and *Stm3*, creating a variable with six possible values. *Info_Set2* is a variable that combines Y and the *Stm3* variable of both committee members in a committee, with 18 possible values. $Asessement$ is the assessment of an evaluator.

In all treatments, a member's ability as determined by the computer has a similar level of entropy, see column (1). The pair of private signals that members receive significantly reduces the entropy, see column (2). Still, considerable uncertainty

[27] As regression techniques to estimate entropy and related variables are unavailable, we use an estimate based on maximum likelihood to determine the empirical entropy. Since the maximum likelihood estimate of entropy is biased downward even asymptotically, we use a bias correction term known as a Miller-Madow bias correction. See Paninski (2003) for details. In Table A.12 in the Appendix, we report the bootstrapped standard errors of these estimates. Calculations were made using the 'infotheo' package in R of Meyer (2014).

Table 11: Entropy and mutual information of ability given various variables

	Entropy	Mutual information of ability given various variables					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	Ability	Confl. Sign.	Info_set2	Info_set	Y=1	Stm3	Assessment
A-NoStm	0.9407	0.0706	0.0050	0.0050	0.0050	-	0.0198
A-Stm	0.9092	0.1058	0.0433	0.0301	0.0020	0.0292	0.0135
NoA-Stm	0.9160	0.0938	0.0732	0.0588	0.0109	0.0488	0.0270

Notes: Maximum likelihood estimates of the entropy of ability and the mutual information of ability given various variables, in bits. A Miller-Madow bias correction has been applied. Column (1) reports the empirically estimated entropy in the ability parameter. The other columns list the estimated mutual information of ability variable given the respective variables. *Confl. Sign.* is a dummy set to 1 if the committee received conflicting signals about the state of nature. *Y=1* is a dummy set to 1 if this committee has taken the decision $Y = 1$. *Stm3* codes the three levels of statements we use {'Low,' 'Confident,' 'Very Confident'}, where 'Low' combines 'Neutral,' 'Doubtful' and 'Very Doubtful.' *Info_Set* combines the information in Y and *Stm3* in a single categorical variable with 2×3 categories. *Info_Set2* combines the information in Y and the *Stm3* variables of both committee members in a single categorical variable with $2 \times 3 \times 3$ categories. *Assessment* is the assessment given by evaluators, transformed to a discrete variable by binning the assessments in 1 percentage-point bins. Since subjects chose not to use decimal places, this is without loss of generality. Table A.12 in the Appendix reports the bootstrapped standard errors of these estimates.

remains. Of the information that is available to the committees via their signals, only a small part is revealed by their choices, see column (3). The mutual information of ability given *Info_set2* or *Info_set* is largest in the absence of a concern with assessments, 0.0732 and 0.0588 bits, respectively. It is only about half that size in the *A-Stm* treatment, 0.0433 and 0.0301 bits, respectively, and more than 10 times smaller in the *A-NoStm* treatment, 0.0050 bits. That is, more of the information about ability is revealed by committee members' when they can use cheap-talk statements and when members don't care about their assessments. A comparison of columns (5) and (6) shows that cheap-talk statements contain considerably more information than the decision, a costly signal: nearly 5 times more in the absence of a concern with assessments and nearly 15 times more in the presence of such concerns.

Of the three treatments, assessments in the *NoA-Stm* treatment contain the most information about members' ability and the least in the *A-Stm* treatment, see column (7). The link between assessments and ability is weak since only a limited amount of information is available to the committee members (compare columns (1) and (2)) and even that information is largely concealed by the behavior of subjects (compare columns (2) and (3)). The measured mutual information in the *A-NoStm* treatment also suggests that randomness has a role to play. In this treatment, the

assessments appear to contain more information about ability than the decision variable, while the decision is the only information the evaluators have to update their beliefs.

When we combine the outcomes of the orthogonality test with the measurement of entropy the following picture arises. When committee members care about their assessments they obfuscate the signals they receive, complicating considerably the formation of assessments.[28] This is reflected in the relatively small amount of information about ability that is present in the assessments in the treatments with a concern with assessments. Whether concerns for assessments are present or not, cheap-talk statements contain considerably more information about a member's ability than the decision the committee takes. In this experiment, words speak louder than actions. This justifies the dependence of assessments on these statements.

7 Conclusion

In this paper, we present the findings of an experiment in which committees of decision makers interact with evaluators who assess them. Evaluators face a difficult problem as they don't observe the state when assessing committee members. The theoretical predictions require a high degree of strategic sophistication. The controlled lab environment allows us to discern whether and why committee members take a decision that looks good but is bad and to measure the quality and determinants of the assessments of evaluators. The chat within the committee sheds light on, *e.g.*, the beliefs about the relationship between assessments and actions taken, and whether the statements sent to evaluators are coordinated or not.

As predicted by theory, in all treatments evaluators assess the ability of committee members higher when committees implement the project than when they maintain the status quo. Once one controls for cheap-talk statements, it becomes clear that evaluators pay considerable attention to the cheap-talk statements of committee members. The orthogonality tests show that this is justified: evaluators use the available information quite efficiently, even when members act strategically to shape the assessments they receive. Unsurprisingly, evaluators struggle to interpret infrequent statements.

The chat proves useful in analysing committee behavior. It shows that reputation concerns induce committee members to discuss the relationship between assessments

[28] Obfuscation caused by a concern with assessments also came to the fore in Table 6: a concern with assessments makes the share of conflicting signals for which a message is used more equal across messages.

and their actions. In fact, when they care about their assessments, a large number of committees relates the assessments they receive to the statements they make, rather than to the decision they take.

Moreover, 30% of members use a statement strategy from which evaluators can't infer anything about their ability. The majority of members, though, condition what they say on their private signals, making it possible for evaluators to glean information about ability from the statements they observe. Evaluators pick this up and make their assessments dependent on the statements they observe. Interestingly, because of their focus on statements as a means to influence assessments, decision making by reputation-concerned committees improves compared to decision making in the absence of cheap-talk statements.

The information-theoretic analysis shows that reputation concerns greatly reduce the amount of information about ability embedded in the committee's observable actions. As a result, the task of evaluators becomes considerably more complicated, especially if they must rely exclusively on the project decision for their assessments. This analysis also shows that in this experiment, on average, words speak louder than costly actions.

Our finding that decision-making by reputation-concerned agents improves if they can use cheap-talk statements suggests that if real-world decision makers care about being assessed as well-informed, one should require their decisions to be accompanied by statements. Our experiment also suggest that it may be hard to have both undistorted decisions and statements that accurately reflect the confidence decision makers have in their decisions. The heterogeneity of strategies used by our subjects implies that careful selection can alleviate the tension as there are subjects who don't distort the decision and make their statements dependent on the private signals they receive.

In reality, evaluators may have different incentives than providing accurate assessments. One possibility that should not be excluded is that they care about coming across as capable or well-informed themselves. The implications this has for their assessments of these committees and, in turn, for the behavior of the committees they evaluate is an interesting area for future research.

Our experimental approach suggests a strategy for acquainting subjects with a full game. Instead of using repetition as a tool, decomposition of the full game might be another option. Rather than letting subjects experience the full-fledged version of the game from the start of the experiment, a treatment can consist of various rounds of a simple version, to which additional layers of complexity are being added. Subjects - both committee members and evaluators - can first gain

experience with a situation in which committee members don't care about reputation concerns. This simplifies the environment of both types of subjects. In a second step, reputation concerns are added. It would be interesting to see whether the two ways of stimulating learning – repetition or decomposition – yield different outcomes.

References

Berg, Joyce E, John W Dickhaut, and Chandra Kanodia (2009) 'The role of information asymmetry in escalation phenomena: Empirical evidence.' *Journal of Economic Behavior & Organization* 69(2), 135–147

Cai, Hongbin, and Joseph Tao-Yi Wang (2006) 'Overcommunication in strategic information transmission games.' *Games and Economic Behavior* 56(1), 7–36

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *The Review of Economics and Statistics* 90(3), 414–427

_ (2011) 'Robust inference with multiway clustering.' *Journal of Business & Economic Statistics* 29(2), 238–249

Dewatripont, Mathias, Ian Jewitt, and Jean Tirole (1999a) 'The economics of career concerns, part 1: comparing information structures.' *Review of Economic Studies* 66(1), 183–198

_ (1999b) 'The economics of career concerns, part 2: application to missions and accountability of government agencies.' *Review of Economic Studies* 66(1), 199–2017

Dickhaut, John W, Kevin A McCabe, and Arijit Mukherji (1995) 'An experimental study of strategic information transmission.' *Economic Theory* 6, 389–403

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner (2011) 'Individual risk attitudes: Measurement, determinants, and behavioral consequences.' *Journal of the European Economic Association* 9(3), 522–550

Donkers, Bas, Bertrand Melenberg, and Arthur Van Soest (2001) 'Estimating risk attitudes using lotteries: A large sample approach.' *Journal of Risk and uncertainty* 22(2), 165–195

Fama, Eugene F (1980) 'Agency problems and the theory of the firm.' *The Journal of Political Economy* pp. 288–307

Fehrler, Sebastian, and Niall Hughes (2018) 'How transparency kills information aggregation: Theory and experiment.' *American Economic Journal: Microeconomics* 10(1), 181–209

Goeree, Jacob K, and Leeat Yariv (2011) 'An experimental study of collective deliberation.' *Econometrica* 79(3), 893–921

Greiner, Ben (2004) 'An online recruitment system for economic experiments.' *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht* pp. 79–93

Hermalin, Benjamin E, and Michael S Weisbach (2017) 'Assessing managerial ability: implications for corporate governance.' Technical Report

Holmström, Bengt (1999) 'Managerial incentive problems: A dynamic perspective.' *The Review of Economic Studies* 66(1), 169–182

Hossain, Tanjim, and Ryo Okui (2013) 'The binarized scoring rule.' *The Review of Economic Studies* 80(3), 984–1001

Irlenbusch, Bernd, and Dirk Sliwka (2006) 'Career concerns in a simple experimental labour market.' *European Economic Review* 50(1), 147–170

Katok, Elena, and Enno Siemsen (2011) 'Why genius leads to adversity: Experimental evidence on the reputational effects of task difficulty choices.' *Management Science* 57(6), 1042–1054

Keane, Michael P, and David E Runkle (1990) 'Testing the rationality of price forecasts: New evidence from panel data.' *The American Economic Review* 80(4), 714–735

_ (1998) 'Are financial analysts' forecasts of corporate profits rational?' *Journal of Political Economy* 106(4), 768–805

Koch, Alexander K, Albrecht Morgenstern, and Philippe Raab (2009) 'Career concerns incentives: An experimental test.' *Journal of Economic Behavior & Organization* 72(1), 571–588

Levy, Gilat (2007) 'Decision making in committees: Transparency, reputation, and voting rules.' *The American Economic Review* pp. 150–168

Luenberger, David G. (2006) *Information Science* (Princeton University Press)

Mattozzi, Andrea, and Marcos Y. Nakaguma (2017) 'Public versus secret voting in committees.' *working paper*

Meloso, Debrah, Salvatore Nunnari, and Marco Ottaviani (2017) 'Looking into crystal balls: a laboratory experiment on reputational cheap talk.' *working paper*

Meyer, Patrick E. (2014) *infotheo: Information-Theoretic Measures.* R package version 1.2.0

Ottaviani, Marco, and Peter Sørensen (2001) 'Information aggregation in debate: who should speak first?' *Journal of Public Economics* 81(3), 393–421

_ (2006a) 'Professional advice.' *Journal of Economic Theory* 126(1), 120–142

_ (2006b) 'The strategy of professional forecasting.' *Journal of Financial Economics* 81(2), 441–466

Paninski, Liam (2003) 'Estimation of entropy and mutual information.' *Neural computation* 15(6), 1191–1253

Prendergast, Canice, and Lars Stole (1996) 'Impetuous youngster and jaded old-timers: acquiring a reputation for learning.' *Journal of Political Economy* 104(6), 1105–1134

Scharfstein, David S, and Jeremy C Stein (1990) 'Herd behavior and investment.' *The American Economic Review* pp. 465–479

Schlag, Karl H, and Joel J van der Weele (2013) 'Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality.' *Theoretical Economics Letters* 3, 38–142

Shannon, Claude E. (1948) 'A mathematical theory of communication.' *Bell System Technical Journal* 27(3), 379–423

Swank, Otto H, and Bauke Visser (2013) 'Is transparency to no avail?' *The Scandinavian Journal of Economics* 115(4), 967–994

Visser, Bauke, and Otto H Swank (2007) 'On committees of experts.' *The Quarterly Journal of Economics* 122(1), 337–372

A Appendix

A.1 The stochastic scoring rule used to incentivize evaluators

We incentivized evaluators to report their true assessments by rewarding them using a stochastic scoring rule, as in Hossain and Okui (2013) and Schlag and van der Weele (2013).[29] For each assessment, an evaluator received tickets in two lotteries, the H-lottery and the L-lottery. The H-lottery was played if the evaluated member had received high quality information, while the L-lottery was played if he had received low quality information. By increasing the reported assessment, the evaluator increases the number of H-tickets she receives, and reduces the number of L-tickets. An evaluator thus faces a trade-off between the two types of lottery tickets. Winning a lottery yields 40, while loosing a lottery yields 0. In this elicitation, it is incentive compatible for an evaluator to truthfully report the subjective probability she attaches to a member being well-informed, irrespective of the evaluator's risk-preferences. It suffices that she rather has a binary lottery that assigns the higher probability to the larger reward than a binary lottery that assigns the lower probability to the larger reward. Evaluators could insert mock assessments in a box on their screens to observe the number of L-tickets and H-tickets that would result. Moreover, at every desk there was a table indicating for multiples of 5% the corresponding numbers of lottery tickets with either label.

A.2 Personal characteristics

In this section, we show that personal characteristics influence observed behavior to a small extent and leave the role played by the main determinants unaffected. We begin by looking at committee members.

A.2.1 Committee members

Committee members differ in their inclination to vote for $Y = 1$ in case of conflicting signals only, see Figures 4 and 6. We thus select those observations where committees received conflicting signals and run a logit regression of a vote v^1 in favor of $Y = 1$. The results are reported in Table A.1, both without personal characteristics (the odd-numbered columns) and with (the even-numbered columns). *Signal* is a dummy set to 1 if the signal received by this member indicates a good state of

[29] We are grateful to Marco Ottaviani for making available the instructions of the experiment described in Meloso et al. (2017). We used their operationalization of the scoring rule.

nature. As all observations have conflicting signals, the coefficient on *Signal* picks up any excess weight members place on their own positive signal compared to the signal of their group member. In two out of three treatments there is evidence that committee member slightly over-weigh their own signal relative to the signal of his fellow member in deciding how to vote. Importantly, the inclusion of personal characteristics leaves this unchanged.

When it comes to the influence of personal characteristics, we find that a member's self-assessment of his willingness to take risk positively affects his inclination to vote favorably in all treatments. Voting for $Y = 1$ in case of conflicting signals amounts to favoring the decision with uncertain and negative expected project payoff over the decision with zero project payoff for sure. Personal risk tolerance therefore has the expected effect. Age consistently and negatively affects the inclination to vote favorably. This is quite surprising given the limited variance in age. One interpretation is that the slightly older participants are slightly less likely to take a risky decision. This is consistent with findings about the relationship between age and risk taking in other papers, see *e.g.*, Donkers et al. (2001). Studying economics or business seems to decrease the likelihood of voting for $Y = 1$ in two treatments, but not in the *A-Stm* treatment. Gender effects don't seem to play a role, while the year of study shows inconsistent effects across treatments.

We do a similar exercise for the statements made but then over all periods using ordered logistical regressions. A comparison of the odd and even columns in Table A.2 shows that the inclusion of personal characteristics leave the role of the committee's decision and their signals qualitatively unaffected, while none of the demographic variables have a consistent effect on the statements made over the treatments.[30]

A.2.2 Evaluators

Table A.3 reproduces the regressions of Table 4, but with personal characteristics variables replacing the evaluators' FEs. As with the committee members, the main conclusions as to the determinants of assessments are unaffected. Three variables, *Econ_bg*, *Year* and *Age*, are each significant in one treatment, but no systematic effects are found. Unreported regressions reproducing the orthogonality tests, Table 9, show no statistically significant effect of any of the background variables.

[30] Untabulated results for the *Stmt3* variable that groups the low confidence statements yield the same conclusion.

Table A.1: Voting behavior as a function of demographic background

VARIABLES	(1) v^1	(2) v^1	(3) v^1	(4) v^1	(5) v^1	(6) v^1
Signal	0.354	0.470	1.015***	1.195***	0.473**	0.562**
	(0.345)	(0.400)	(0.312)	(0.345)	(0.219)	(0.240)
Age		-0.499***		-0.183**		-0.0855***
		(0.168)		(0.0875)		(0.0280)
Male		-0.570		0.215		0.199
		(0.443)		(0.392)		(0.273)
Econ_bg		-1.696***		0.525		-2.074***
		(0.546)		(0.534)		(0.436)
Year		0.129		0.319**		-0.172*
		(0.171)		(0.155)		(0.0895)
Risk_tolerance		0.363***		0.427***		0.284***
		(0.0911)		(0.104)		(0.0487)
Constant	-0.526**	8.844***	-1.099***	-1.408	-0.676***	2.008**
	(0.247)	(3.351)	(0.236)	(1.598)	(0.159)	(0.864)
Sample	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.
Observations	140	140	192	192	356	356
Subject FE	NO	NO	NO	NO	NO	NO
Period FE	NO	NO	NO	NO	NO	NO
Treatment	A-NoStm	A-NoStm	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: *Signal* is a dummy set to 1 if the signal received by this subject indicated a good state of nature. *Age* is a subject's age in years. *Male* and *Econ_bg* are dummies set to 1 if the subject is male and studies Economics or Business, respectively. *Year* captures the year of study of the subject, running from 1 for bachelor 1, to 6 for master students. *Risk_tolerance* is a subject's self-reported risk tolerance on an 11-point scale running from 1, 'not willing to take any risk,' to 11, 'very willing to take risks.' Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.3 Evaluators treat decision and statements as separate channels of information

In section 4.1, we didn't include interaction terms in the assessment regression. However, to form an assessment, evaluators can use any combination of observable behavior of the committee members they have. In Table A.4, we explicitly allow for possible interactions between decision and statement in the determination of assessments. We find no significant interactions between the statement variables and the decision. This indicates that evaluators treat the channels of information as separate.

Table A.2: Statements as a function of demographic background

VARIABLES	(1) Statement	(2) Statement	(3) Statement	(4) Statement
Y=1	0.378	0.405	-1.452***	-1.550***
	(0.286)	(0.295)	(0.218)	(0.229)
Signal	0.243	0.263	0.866***	0.935***
	(0.223)	(0.228)	(0.162)	(0.166)
Signal_other	-0.0663	-0.101	0.889***	0.948***
	(0.223)	(0.226)	(0.165)	(0.170)
Confl. Signals	-1.634***	-1.661***	-3.531***	-3.590***
	(0.194)	(0.199)	(0.174)	(0.176)
Age		-0.0585		0.0654***
		(0.0395)		(0.0141)
Male		-0.234		0.148
		(0.196)		(0.135)
Econ_bg		-0.0511		0.289
		(0.284)		(0.213)
Year		0.273***		-0.111**
		(0.0821)		(0.0470)
Risk_tolerance		0.000610		0.0408*
		(0.0464)		(0.0228)
/cut1	-7.090***	-7.702***	-6.904***	-5.377***
	(1.012)	(1.239)	(0.352)	(0.549)
/cut2	-3.897***	-4.504***	-4.187***	-2.626***
	(0.260)	(0.758)	(0.190)	(0.470)
/cut3	-2.234***	-2.821***	-2.207***	-0.617
	(0.178)	(0.731)	(0.139)	(0.458)
/cut4	-1.152***	-1.718**	1.005***	2.656***
	(0.155)	(0.724)	(0.110)	(0.458)
Observations	660	660	1,140	1,140
Subject FE	NO	NO	NO	NO
Period FE	NO	NO	NO	NO
Treatment	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: *Signal* is a dummy set to 1 if the signal received by this subject indicated a good state of nature. *Signal_other* is a dummy set to 1 if the signal received by this subject's fellow committee member indicated a good state of nature. *Confl. Signals* is a dummy set to 1 if the signals received by this subject and his fellow committee member indicated two different states of nature. *Age* is a subject's age in years. *Male* and *Econ_bg* are dummies that equal 1 if the subject is male, and studies Economics or Business respectively. *Year* captures the year of study of the subject, running from 1 for bachelor 1, to 6 for master students. *Risk_tolerance* is a subject's self-reported risk tolerance on an 11-point scale running from 1, 'not willing to take any risk,' to 11, 'very willing to take risks.' Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.4 Dynamic effects I: lagged variables

Since the subjects in our experiments play multiple rounds we check whether there are dynamic effects or time trends that affect our conclusions on committee behav-

Table A.3: Evaluators' behavior as a function of demographic background

VARIABLES	(1) Assessment	(2) Assessment	(3) Assessment	(4) Assessment	(5) Assessment	(6) Assessment
Yellow	9.834***	10.06***	2.701	2.568	5.667***	5.594***
	(1.631)	(1.894)	(1.763)	(1.743)	(1.460)	(1.461)
Very_Confident			17.10***	16.82***	25.91***	26.11***
			(3.480)	(3.501)	(3.050)	(2.974)
Confident			11.19***	11.31***	16.33***	16.07***
			(3.257)	(3.464)	(2.346)	(2.317)
Same_statement			0.609	0.148	2.005**	2.277***
			(1.724)	(1.491)	(0.751)	(0.729)
Age		1.070		-0.908*		0.224
		(0.534)		(0.406)		(0.463)
Male		3.221		0.735		-3.246
		(2.973)		(2.060)		(2.369)
Econ_bg		0.754		0.0262		5.348**
		(3.024)		(3.090)		(2.474)
Year		2.066**		0.277		0.197
		(0.645)		(1.011)		(0.966)
Risk_tolerance		1.055		0.580		-0.393
		(1.344)		(0.608)		(0.667)
Constant	55.40***	19.14	53.08***	68.59***	46.77***	41.22***
	(1.159)	(13.14)	(3.864)	(9.914)	(2.221)	(11.55)
Observations	1,680	1,680	2,640	2,640	5,040	5,040
R^2	0.106	0.248	0.162	0.178	0.383	0.399
Cl-level	match	match	match	match	match	match
Clusters	6	6	10	10	21	21
Subject FE	NO	NO	NO	NO	NO	NO
Period FE	YES	YES	YES	YES	YES	YES
Treatment	A-NoStm	A-NoStm	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: $Y = 1$ is a dummy set to 1 if this committee chooses $Y = 1$, *Very Confident* and *Confident* are dummies set to 1 if the corresponding statement is received. *Same_Statement* is a dummy set to 1 if both members of this committee used the same statement. *Age* is a subject's age in years. *Male* and *Econ_bg* are dummies that equal 1 if the subject is male, and studies Economics or Business, respectively. *Year* captures the year of study of the subject, running from 1 for bachelor 1 to 6 for master students. *Risk_tolerance* is a subject's self-reported risk tolerance on an 11-point scale running from 1, 'not willing to take any risk,' to 11, 'very willing to take risks.' Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

ior and assessments of evaluators. In this section, we study the effects of lagged variables. In section A.5, we compare behavior in the first half and the second half of the experiment to investigate any learning effects. We find that, although lagged variables play some role in all treatments and both for committee members and evaluators, they leave the relations between current-period variables unaffected. Similarly, even though we find some small differences in the size of the coefficients, the relations identified in the main text persist in both the first and second half of

Table A.4: Formation of assessment, full strategy

VARIABLES	(1) Assessment	(2) Assessment
$Y = 1$	2.675	0.865
	(2.709)	(2.064)
Very Confident	17.36***	25.29***
	(2.328)	(3.300)
Very Confident $\times$ $Y = 1$	-1.132	4.367
	(3.822)	(3.428)
Confident	12.97***	15.25***
	(2.522)	(2.655)
Confident $\times$ $Y = 1$	-1.335	4.136
	(4.403)	(2.815)
Same Statement	2.164*	1.162
	(0.983)	(1.012)
Same Statement $\times$ $Y = 1$	0.859	0.877
	(2.971)	(1.316)
Constant	51.93***	44.75***
	(3.251)	(2.140)
Observations	2,640	5,040
R^2	0.407	0.572
Cl-level	match	match
Clusters	10	21
Subject FE	YES	YES
Period FE	YES	YES
Treatment	A-Stm	NoA-Stm

Notes: *Assessment* is the assessment given by an evaluator for a particular member, on the original 100-point scale. $Y = 1$ is a dummy set to 1 if this member's committee chooses $Y = 1$. *Very Confident* and *Confident* are dummies set to 1 if the member uses that cheap-talk statement. *Same Statement* is a dummy that is set to one if this member uses the same cheap-talk statement as his fellow committee member in that period. Fixed effect specification. Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

the experiment. Given the higher number of periods, it is particularly reassuring that differences are small in the *NoA-Stm* treatment.

A.4.1 Committee members

Table A.5: Relationship between lagged variables and current voting behavior

VARIABLES	(1) v^1	(2) v^1	(3) v^1	(4) v^1	(5) v^1	(6) v^1
Signal	0.354	0.297	1.015***	1.351***	0.473**	0.501**
	(0.345)	(0.368)	(0.312)	(0.365)	(0.219)	(0.227)
Lag_Y=1		6.402*		7.436***		14.86
		(3.361)		(2.837)		(713.4)
Lag_Very_confident				-1.755**		-0.191
				(0.758)		(0.341)
Lag_Confident				-2.397***		-0.397
				(0.750)		(0.293)
Lag_Same_statement				-0.974**		0.216
				(0.440)		(0.250)
Lag_Ability		-0.109		0.00927		-0.0208
		(0.413)		(0.404)		(0.247)
Lag_Score		0.0421		0.0720***		0.113
		(0.0303)		(0.0245)		(5.945)
Lag_State_of_world		0.231		-0.819		-0.296
		(0.591)		(0.650)		(0.416)
Lag_Succes		-11.76*		-15.35***		-26.90
		(6.954)		(5.470)		(1,367)
Constant	-0.526**	-2.914*	-1.099***	-3.657***	-0.676***	-0.627**
	(0.247)	(1.681)	(0.236)	(1.373)	(0.159)	(0.315)
Sample	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.
Observations	140	134	192	178	356	344
Subject FE	NO	NO	NO	NO	NO	NO
Period FE	NO	NO	NO	NO	NO	NO
Treatment	A-NoStm	A-NoStm	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: There is one current-period variable: *Signal* is a dummy set to 1 if the corresponding signal equals s^g. All variables that start with *lag_* are measured in the period before the conflicting signals occurred. Lag_Y=1 is set to 1 if this committee chose Y=1 in the last round, *Lag_Very_confident* and *Lag_Confident* keep track of the statements used in the last period. *Lag_Ability* is a dummy that equals 1 if this committee member had high ability, i.e. got information from an H-box, last period. *Lag_Score* measures the points earned by a member in the last period. *lag_State_of_world* is a dummy that equals 1 if the state of the world in the last period was good ($\mu = h$). *lag_Success* is dummy that equals 1 if the committee chose $Y = 1$ and $\mu = h$ (*i.e.*, it is the interaction between *lag_State_of_world* and *lag_*$Y = 1$). Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.5 looks at voting behavior in rounds in which members of a committee receive conflicting signals. As far as current-period signals are concerned, we find again that members tend to overweigh their own private signal. As for the lagged variables, the findings suggest that the direction of the effect of each of them is

Table A.6: Relationship between lagged variables and current statements

VARIABLES	(1) Stmt3	(2) Stmt3	(3) Stmt3	(4) Stmt3
Y=1	1.491***	1.983***	0.370*	0.491**
	(0.297)	(0.340)	(0.222)	(0.233)
Signal	-0.285	-0.593**	-0.0679	-0.159
	(0.230)	(0.261)	(0.170)	(0.178)
Signal_other	-0.676***	-0.965***	-0.0137	-0.0162
	(0.227)	(0.260)	(0.168)	(0.176)
Lag_Y=1		0.219		16.34
		(1.397)		(553.6)
Lag_Very_confident		1.636***		1.493***
		(0.359)		(0.197)
Lag_Confident		0.519		-0.0113
		(0.324)		(0.155)
Lag_Ability		-0.0230		0.00404
		(0.214)		(0.130)
Lag_Score		0.00190		0.133
		(0.0117)		(4.613)
Lag_Same_statement		1.055***		0.0600
		(0.216)		(0.128)
Lag__State_of_world		0.552*		0.400*
		(0.316)		(0.209)
Lag_Success		-0.954		-31.28
		(2.697)		(1,061)
/cut1	-1.690***	-0.114	-0.855***	-0.542***
	(0.152)	(0.661)	(0.0943)	(0.184)
/cut2	-0.706***	1.070	1.424***	1.895***
	(0.132)	(0.660)	(0.102)	(0.194)
Observations	660	616	1,140	1,102
Subject FE	NO	NO	NO	NO
Period FE	NO	NO	NO	NO
Treatment	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: There are three current-period variables: $Y = 1$ is a dummy that equals 1 if the committee chooses $Y = 1$; *Signal* and *Signal_other* are dummies set to 1 if the corresponding signal equals s^g. All variables that start with *lag_* are measured in the period before the period under analysis. Lag_Y=1 is set to 1 if this committee chose Y=1 in the last round, *Lag_Very_confident* and *Lag_Confident* keep track of the statements used in the last period. *Lag_Ability* is a dummy that equals 1 if this committee member had high ability, i.e. got information from an H-box, last period. *Lag_Score* measures the points earned by a member in the last period. *Lag_same_statement* is set to 1 if this committee member made the same statement as his fellow committee member last period. *Lag_State_of_world* is a dummy that equals 1 if the state of the world in the last period was good ($\mu = h$). *lag_Success* is dummy that equals 1 if the committee chose $Y = 1$ and $\mu = h$ (*i.e.*, it is the interaction between *lag_State_of_world* and *lag_Y* $= 1$). All analyses are Ordered Logit regressions. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

the same across the three treatments, and a concern with assessments makes some of those effects statistically significant. In particular, the previous decision and statement appear to have some impact on voting in the current period in the *A*-

Stm treatment. Comparing the coefficients on the dummies for $lag\_Y = 1$ and *lag_Success*, we see that members in committee that chose $Y = 1$ in the last period are more likely to vote for $Y = 1$, but only if it was the incorrect decision. This is consistent with a belief in the law of small numbers. The coefficient of *lag_Score* shows that members with a relatively high score in the last round are more likely to vote for $Y = 1$, particularly in the *A-Stm* treatment where the score depends most on the statements made. This effect is, however, relatively weak. In the *A-Stm* treatment, past statements also have some role to play. Committee members who expressed confidence in the last period are less likely to vote v^1 in case of conflicting signals in the current round. Furthermore, untabulated regressions show that in all treatments, the signals almost perfectly predict the votes if members receive the same signals also when including lagged variables.

Table A.6 turns to statements, reporting ordered logistic regressions against three levels of confidence stated by committee members. As before, we group the observations of members stating 'Very Doubtful,' 'Doubtful,' and 'Neutral' into one category. The direct comparison between columns (1) and (2), and between (3) and (4), show that the coefficients on current variables hardly change when including lagged variables, as with voting behavior. We find some effects of previous statements on current statements. A committee member who chose 'Very Confident' in one period is more likely to choose 'Very Confident' in the next period. Furthermore, members who had matching statements in the last period are more likely to make high statements in the *A-Stm* treatment, possibly through ceiling effects. However, there does not seem to be a strong effect of the outcomes and choices from the previous period on the relation found between current-period variables of interest.

A.4.2 Evaluators

Table A.7 compares the formation of assessments, with and without lagged variables. A comparison of the columns (1) with (2), (3) with (4), and (5) with (6), one sees that adding lagged variables does not change the relation between contemporaneous variables in any meaningful way in any of the treatments. Recall that each evaluator assesses both members of two committees. Choices made by these members are presented on four positions on an evaluator's screen. The position where the choices of a member are presented varies randomly across rounds. All lagged variables in table A.7 are measured from the point of view of the evaluator, so that *Lag_Assessment* is the assessment this evaluator gave last period to the committee member on the same position on the screen. There appears to be a small positive correlation between the assessments made on a specific position on an evaluator's

Table A.7: Relationship between lagged variables and current assessments

VARIABLES	(1) Assessment	(2) Assessment	(3) Assessment	(4) Assessment	(5) Assessment	(6) Assessment
Y=1	9.602***	9.511***	2.222	2.452	4.631***	4.512***
	(1.605)	(1.638)	(1.874)	(1.844)	(1.318)	(1.326)
Very_Confident			17.07***	16.11***	26.97***	26.80***
			(2.614)	(2.301)	(2.866)	(2.870)
Confident			12.57***	11.40***	16.61***	16.49***
			(2.669)	(2.685)	(2.371)	(2.370)
Same_statement			2.466**	2.254*	1.610*	1.887**
			(0.760)	(1.059)	(0.870)	(0.888)
Lag_Yellow		0.0440		-0.485		-0.561
		(0.471)		(0.639)		(0.510)
Lag_Assessment		0.0655		0.172		0.0676***
		(0.0422)		(0.0982)		(0.0220)
Lag_Ability		-0.605		1.048		0.341
		(0.316)		(0.947)		(0.397)
Lag_Score		0.0149		-0.00500		0.00766
		(0.00943)		(0.00637)		(0.00705)
Lag_Very_Confident				-4.050*		-2.217**
				(1.942)		(0.946)
Lag_Confident				-2.294		-1.132
				(1.661)		(0.721)
Lag_Same_statement				-2.402***		-0.483
				(0.682)		(0.314)
Constant	54.04***	47.60***	51.98***	46.19***	43.38***	40.86***
	(1.463)	(2.803)	(3.329)	(5.807)	(2.077)	(2.659)
Observations	1,680	1,568	2,640	2,464	5,040	4,872
R^2	0.518	0.536	0.407	0.446	0.569	0.578
Cl-level	match	match	match	match	match	match
Clusters	6	6	10	10	21	21
Subject FE	YES	YES	YES	YES	YES	YES
Period FE	YES	YES	YES	YES	YES	YES
Treatment	A-NoStm	A-NoStm	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: *Assessment* is the assessment given by an evaluator for a particular member, on the original 100-point scale. $Y = 1$ is a dummy set to 1 if this members' committee chooses $Y = 1$. *Very Confident* and *Confident* are dummies that are set to 1 if the member uses that cheap-talk statement. *Same Statement* is a dummy that is set to 1 if this member uses the same cheap-talk statement as his fellow committee member in that period. All variables that start with "Lag_" are measured in the period before, on the same, randomly assigned spot on the screen of the evaluator. So, *Lag_Y=1* is a dummy that is equal to 1 if the dummy $Y = 1$ was equal to 1 in the previous period in the same of the four spots on the evaluator's screen. Last periods assessments in the original 100-point scale is captured in *Lag_Assessment*. The dummies *Lag_Confident*, *Lag_Very Confident*, *Lag_Same Statement*, and *Lag_Ability* are dummies set to 1 if the corresponding dummies were equal to 1 in the last period. *Lag_Score* measures the points earned by this evaluator in the last period. Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

screen in the *NoA-Stm* treatment. However, this effect is very small: an increase in the past assessment by one standard deviation (17.6 points) has less effect than the difference between committees with and without *Same Statement*. The most significant effects appear from *Lag_Same Statement* and *Lag_Very Confident*. However,

their effect on assessments is dwarfed by the effect of contemporaneous statements.

A.5 Dynamic effects II: first versus second half of the experiment

Dynamic effects that build up slowly over rounds cannot always be detected by the introduction of lagged variables. To find any evidence of, *e.g.* learning and fatigue, we compare behavior of our subjects between the first and the second half of the experiment. We generally find little difference between the two halves of the experiment.

A.5.1 Committee Members

Committee members make one or two choices in each period, what to vote and, depending on the treatment, what to state. In section 4, we saw that if committee members receive the same signal, they follow their signals. The relevant part of the committee members' strategy is how they vote in case of conflicting signals. Table A.8 compares voting behavior across the two halves of the experiment. It suggests that voting behavior is stable across the two halves in each treatment. Committee members are more likely to vote for $Y = 1$ if they receive a positive signal, indicating overweighing of their own signal in both halves of the experiment. The constants and coefficients have the same sign and are of comparable size across halves in all treatments. Due to the smaller number of observations per regression, significance levels do change a bit.

A two-sample test of proportions does not suggest there is any differences in the likelihood of voting for $Y = 1$ in periods with conflicting signals across the two halves of the experiment (two-sided p-values larger than 0.25 in all treatments).

If we look at the statements made by the committee members, a similar pattern occurs. Table A.9 shows that coefficient sizes are very similar in the *A-Stm* treatment across the two halves. A small difference occurs in the *NoA-Stm* treatment where the coefficient on *Signal* and *Signal_other* change sign. However, these are insignificant. Since the underlying variables are positively correlated they are difficult to estimate in any case. If we look at the average statement given in both treatments, we see small non-significant (two-sided t-test unequal variance, p-values above 0.1) differences between the first and second half.

Table A.8: Voting with conflicting signals, first and second half of the experiment

VARIABLES	(1) v^1	(2) v^1	(3) v^1	(4) v^1	(5) v^1	(6) v^1
Signal	0.323	0.637**	0.903**	1.115***	0.501	0.225
	(0.305)	(0.316)	(0.457)	(0.428)	(0.504)	(0.475)
Constant	-0.481**	-0.885***	-0.949***	-1.229***	-0.738**	-0.336
	(0.218)	(0.233)	(0.340)	(0.328)	(0.367)	(0.338)
Sample	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.	Confl. sign.
Periods	3-17	18-32	3-9	10-18	3-9	10-18
Observations	178	178	86	106	68	72
Clusters	NO	NO	NO	NO	NO	NO
Subject FE	YES	YES	YES	YES	YES	YES
Period FE	NO	NO	NO	NO	NO	NO
Treatment	NoA-Stm	NoA-Stm	A-Stm	A-Stm	A-NoStm	A-NoStm

Notes: v^1 is a dummy that equals 1 if this committee member votes for $Y = 1$ this period. *Signal* is a dummy equal to 1 if this committee member received a signal s^g. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.9: Statements, first and second half of the experiment

VARIABLES	(1) Stmt3	(2) Stmt3	(3) Stmt3	(4) Stmt3
Y=1	1.311***	1.764***	0.0281	0.749**
	(0.420)	(0.419)	(0.308)	(0.323)
Signal	-0.156	-0.459	0.113	-0.275
	(0.332)	(0.319)	(0.236)	(0.246)
Signal_other	-0.532	-0.874***	0.0285	-0.0657
	(0.328)	(0.316)	(0.236)	(0.240)
Periods	3-9	10-17	3-17	18-32
Observations	308	352	570	570
Subject FE	NO	NO	NO	NO
Period FE	NO	No	NO	NO
Treatment	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: *Stmt3* is a categorical variable capturing the statements made by the committee members. It runs from 3 "Very Confident" to 1 "Neutral or below". *Signal* and *Signal_other* are dummies equal to 1 if the corresponding committee member received a signal indicating a good state of the world that period. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.5.2 Evaluators

Table A.10 compares the assessments in the first and second half of a treatment. All coefficients have the same sign and similar sizes across halves in each treatment. Furthermore, they are also comparable to the estimates obtained in Table 4. A direct comparison of averages using t-tests gives a similar result. The difference in average assessment is insignificant for the *A-Stm* treatment, while it is statistically

significant but small in the *NoA-Stm* (diff 2, $p < 0.01$ in a t-test) and the *A-NoStm* treatments (diff 1.6, $p = 0.03$ in a t-test).

Table A.10: Assessments, first and second half of the experiment

VARIABLES	(1) Assessment	(2) Assessment	(3) Assessment	(4) Assessment	(5) Assessment	(6) Assessment
Y=1	10.72***	8.320***	2.654	1.461	5.102***	4.180**
	(1.843)	(1.642)	(2.818)	(1.405)	(1.299)	(1.638)
Very Confident			18.39***	15.35***	28.28***	25.83***
			(2.893)	(2.538)	(3.223)	(2.818)
Confident			12.41***	11.86***	17.56***	15.57***
			(2.813)	(2.949)	(2.443)	(2.539)
Same Statement			2.462	2.389*	1.587*	1.142
			(1.398)	(1.087)	(0.902)	(1.141)
Constant	53.65***	56.91***	51.12***	52.00***	42.47***	46.78***
	(0.946)	(1.458)	(3.480)	(2.362)	(2.065)	(1.982)
Periods	3-9	10-17	3-9	10-17	3-17	18-32
Observations	784	896	1,232	1,408	2,520	2,520
R^2	0.499	0.586	0.393	0.510	0.566	0.604
Cl-level	match	match	match	match	match	match
Clusters	6	6	10	10	21	21
Subject FE	YES	YES	YES	YES	YES	YES
Period FE	YES	YES	YES	YES	YES	YES
Treatment	A-NoStm	A-NoStm	A-Stm	A-Stm	NoA-Stm	NoA-Stm

Notes: *Assessment* is the assessment given by an evaluator for a particular member, on the original 100-point scale. $Y = 1$ is a dummy set to 1 if this member's committee chooses $Y = 1$. *Very Confident* and *Confident* are dummies that are set to 1 if the member uses that cheap-talk statement. *Same Statement* is a dummy set to one if this member uses the same cheap-talk statement as his fellow committee member in that period. Fixed effect specification. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.6 Power orthogonality test

A concern that one could have with the null-results in the orthogonality tests in columns (1)–(3) of Table 9 is that they stem from a lack of power of the orthogonality test rather than from evaluators using information efficiently. We argue that such a concern seems unjustified by showing that our test does have the power to detect unused information. Table A.11 reproduces the regressions of Table 9 and adds a dummy variable *Confl. Signals* that is equal to 1 if committee members receive conflicting signals about the state of nature in a particular round. This dummy captures all information about members' ability that becomes available in the experiment – it is a sufficient statistic for the information about ability contained in the statements and decisions. Conflicting signals is a sure sign that at least one member is of low ability. Hence, *Confl. Signals* should strongly correlate with abil-

Table A.11: Power of orthogonality tests

VARIABLES	(1) Mistake	(2) Mistake	(3) Mistake
Confl. Signals	-30.77***	-37.54***	-34.13***
	(10.13)	(12.31)	(11.04)
Y=1	-5.491*	-6.961**	-1.924
	(3.115)	(2.827)	(3.751)
Very Confident		-5.960	-22.86***
		(8.323)	(7.395)
Confident		0.666	-14.59***
		(1.304e+19)	(5.307)
Same Statement		-1.119	-2.330
		(5.255)	(2.687)
Constant	17.09***	18.02***	27.91***
	(0)	(0)	(0)
Observations	1,680	2,640	5,040
R-squared	0.084	0.109	0.060
Cl-level	Match & Period	Match & Period	Match & Period
Subject FE	NO	NO	NO
Period FE	NO	NO	NO
Treatment	A-NoStm	A-Stm	NoA-Stm

Notes: *Mistake* is equal to the difference between the realized probability of ability and the assessments of this ability both on a 100 point scale. *Confl. Signals* is a dummy set to 1 if this members' committee received conflicting signals. $Y = 1$ is a dummy set to 1 if this members' committee chooses $Y = 1$. *Very Confident* and *Confident* are dummies that are set to 1 if the member uses that cheap-talk statement. *Same Statement* is a dummy that is set to one if this member uses the same cheap-talk statement as his fellow committee member in that period. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

ity. Evaluators don't observe the signals received by the committee members and can base their evaluations only on observed committee behavior – i.e. the decision and, possibly, statements. As a result, the correlation with true ability should be substantially weaker for assessments than for the *Confl. Signals* dummy. As the mistake variable is the difference between true ability and assessments, *Confl. Signals* should be systematically related to the *Mistake* variable. The table shows that the coefficient of *Confl. Signals* is large and strongly significant in all treatments. This means that these regressions have the power to pick up systematic deviations from information efficiency if an unused informative signal is used.

A.7 Standard error entropy estimates

In Table 11, we reported the entropy of ability and the mutual information of ability given any variable that can be observed by some subject in the experiment. Table

A.12 reports the same information and adds bootstrapped standard errors. Bootstrapping was done by drawing random samples (with replacement) of the same size as the original sample. Standard errors are based on 10,000 random samples.

Table A.12: Maximum likelihood estimates and bootstrapped standard errors of entropy and mutual information

	Entropy	Mutual information of ability given various variables					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	Ability	Confl. Sign.	Info_set2	Info_set	Y=1	Stm3	Assessment
A-NoStm	0.9407	0.0706	0.0050	0.0050	0.0050	NA	0.0198
	(0.0100)	(0.0108)	(0.0031)	(0.0031)	(0.0031)	NA	(0.0091)
A-Stm	0.9092	0.1058	0.0433	0.0301	0.0020	0.0292	0.0135
	(0.0097)	(0.0107)	(0.0074)	(0.0061)	(0.0031)	(0.0058)	(0.0061)
NoA-Stm	0.9160	0.0938	0.0732	0.0588	0.0109	0.0488	0.0270
	(0.0066)	(0.0072)	(0.0064)	(0.0058)	(0.0025)	(0.0053)	(0.0049)

Notes: Maximum likelihood estimates of the entropy of ability and the mutual information of ability given various variables, in bits. A Miller-Madow bias correction has been applied. Bootstrapped standard errors based on 10,000 repetitions are in parentheses.
Column (1) reports the empirically estimated entropy in the ability parameter. The other columns list the estimated mutual information of ability variable given the respective variables. *Confl. Sign.* is a dummy set to 1 if the committee received conflicting signals about the state of nature. *Y=1* is a dummy set to 1 if this committee has taken the decision $Y = 1$. *Stm3* codes the three levels of statements we use {'Low,' 'Confident,' 'Very Confident'}, where 'Low' combines 'Neutral,' 'Doubtful' and 'Very Doubtful.' *Info_Set* combines the information in Y and *Stm3* in a single categorical variable with 2×3 categories. *Info_Set2* combines the information in Y and the *Stm3* variables of both committee members in a single categorical variable with $2 \times 3 \times 3$ categories. *Assessment* is the assessment given by evaluators, transformed to a discrete variable by binning the assessments in 1 percentage-point bins. Since subjects chose not to use decimal places, this is without loss of generality.

What is noticeable about the standard errors is their small size. The bootstrapped distributions of entropy and mutual information are quite homogenous and this is reflected in the standard errors. Note that the entropy of a variable is independent of the values that the variable takes one – it only depends on the distribution of its probability mass. Thus, small standard errors mean that the probability distribution of the underlying variables is fairly stable.

A.8 Chatboxes

Each committee member had access to a chatbox that allowed him to freely communicate within his committee. We have analyzed the content of the chatboxes to get more information about among others the thought-process of committee members. This part of the appendix starts by showing some relevant excerpts from these chatboxes in section A.8.1. Section A.8.2 contains the coding instructions given to

the two RAs who coded the content of the chat conversations. The results of the coding are presented in Appendix A.8.3.

A.8.1 Excerpts from the chats

The excerpts that follow – verbatim if the original is in English, translated if in Dutch – present the chronological order in which text lines appeared in the chat box. As a result, one member's line of text may relate to what the other member wrote not in the last line but before. The text of one member is reported in normal script, while the text of the other member is italicized. Chat lines that a member separated by a carriage return are separated by a forward slash. We refer to a committee member by his unique subject_ID number and to a committee by its unique committee_ID number.

When assessments matter to committee members, some committees discuss what drives these assessments. In the majority of cases, this discussion takes place after observing assessments for a number of rounds. We present two excerpts that illustrate this observation. For many rounds, members of committee 13 (*A-Stm*) have been using statements that reflect their signals. In round 11 both report 'Doubtful' after receiving conflicting signals, while in round 12 they report 'Very Confident' after two yellow signals. The chat in round 13 reads as follows.

Excerpt 1. Member 25 : yellow here Member 26: *ok now I got blue / hmm do you wanna risk it again?* I don't, but what about you? *I dont know... btw I noticed that we always get better evaluation when we report high confidence* oh really? let's make use of that then haha / so I'll vote high confidence, but which color.... [...] [After they cast their votes, the chat continues:] so very confident *ok I guess lets put high confidence and see what happens / yes* alright [And at the start of the next round:] *see we got better evaluation* yes I saw, good point :D.

The next excerpt is from session 9, committee 82 (*A-NoStm*). In round 10 they receives conflicting signals. They vote for $Y = 1$ and earn 110 project points (besides assessment points). In the next round, they again receive conflicting signals.

Excerpt 2.[31] Member 163:[32] I say we both do again yellow because I notice that they give us a lot and in the end you lose little / or win a lot Member 164: *Yes, that's right* Especially the last I hope hahah *So the gains > costs if it goes wrong* Yes exactly / Do it *Allright yellow then* [In round 14, they receive conflicting signals. The following chat ensues:] *Blue* [...] yellow now *Well if we vote blue we always get*

[31]This excerpt is translated from Dutch.
[32]This excerpt is translated from Dutch.

around 40 points / If we both do yellow than always around 80 / So the difference is -40 or 40 no 180 right / so -40 or 180 *Yeah, ok true* often *I mean if we vote blue* Oooh *If we do yellow than yes* Yes haha / what does your intuition say *Yes take risk* oke let's do it *Yellow it is*

Few committees take a more cerebral approach to learning about the determinants of assessments, as illustrated by committee 11 (*A-Stm*). In round 1, member 22 suggests to state Confident. After casting their votes in round 2, the following chat develops.

Excerpt 3.[33] Member 21: What I'm thinking / The more enthusiastic we are about our choice, the more points we can get from the evaluators [...] Member 22: *Yes I was thinking the same / One gets more points anyway / Right?* If they give them yeah *Shall we try doubtful / See what happens?* Oh right, test round / okay / let's do it [At the start of round 3, they continue] less points from evaluators / remain enthusiastic *uhu*

The next two excerpts show discussions about the value of a united front.

Excerpt 4. Member 221 in committee 111 (*A-Stm*) in round 4: to maximize our earning. / We should always choose confident/ very confident together. Member 222: *yes!* so that we don't have conflicts and evaluators will give high score to us

Excerpt 5.[34] Member 22 in committee 11 (*A-Stm*) in round 3: *Would it matter if we always go for the same degree?* Member 5: No idea *Whether evaluators assess on that* In the example everybody had a different opinion / but if decision makers do not agree, I would not trust it very much

Excerpt 5 shows that some committee members move beyond simply discussing an observed pattern: they put themselves in the shoes of evaluators to better understand what statements to use. The following three excerpts provide further examples of this capacity.

Excerpt 6.[35] Member 41 in committee 21 (*A-Stm*) in round 13: we must pick high confidence / because they will rate us higher Member 42: *yes* that works out better for us / also if we don't know for sure *every round very confident?* haha chill *okee haha* [They cast their votes and proceed] Every now and then normal / normal confident *yes otherwise it raises suspicion* exactly

Excerpt 7. Member 45 in committee 23 (*A-Stm*) in round 6: lets always say confident or very confident / [...] / so they get tricked into thinking we have a high

[33]This excerpt is translated from Dutch.
[34]This excerpt is translated from Dutch.
[35]This excerpt is translated from Dutch.

chance and we get better pay outs / alright? Member 46: *aye sir / okay* [But in round 10, after presenting their private signals, member 9 comes back to this line of reasoning:] lets choose blue and confident this time / not very / otherwise they will not believe us anymore :p *not very confident?* got points so far *i think its pretty sure we get blue tho xD* good* / haha / TRUE / alright lets go very *just choose confident :P*

Excerpt 8.[36] Member 22 in committee 11 (*A-Stm*) in round 5: *I don't understand why we wouldn't do confident all the time / Or am I overlooking something/ I have got blue by the way* Member 21: I have blue too / If they don't trust us they go lower than if very confident were justified / but they only have us as a source / so they can't do much / so I keep on sending confidents / shall we choose blue by the way? *And 3 of the four follow our degree of confidence* [In round 8, after casting their votes:] what was your average this round? *The payment structure of the evaluators is different* what do you mean? *So why not go for very confident every round / It is their business to make a good assessment of us, right* Yeah yeah that is right *I think about 75 on average* they haven't got anything else to base it on / i had 61 *They don't see our degrees of conf / Indeed / I didn't look what the average was* they do see them / have a look at page 10 —*Where does it say so? / Oooh yeah* but doesn't matter *I see* if they don't trust us because we are overenthusiastic they still don't have anything else *Then we should go for the same every round* simply never do doubtful *What do you mean? / No indeed* whatever if they find out our very confidents don't make any sense / what can they do? *TRUE* nothing / they don't know anything *No indeed / So always ticking the same is the best we can do* so i go and vote *Yes* yes just do good *Very conf*

A.8.2 Scheme used for coding chat

We provided instructions to two research assistants (RAs) on how to code messages in the chatbox. Each research assistant was given a spreadsheet containing all the chat-conversations, with one message per line. All messages were sorted so that the conversations were displayed per committee and in chronological order. Period, member and committee identifiers were also shown. For each message sent, both RAs were instructed to fill out a table with a column for each variable encoded. A mark in a cell indicated that this message contained an instance of the corresponding variable. Tables A.13 and A.14 show the exact explanations given to the RAs on what variables mean and which values it can take on. Furthermore, examples were

[36] This excerpt is translated from Dutch.

provided of messages that should be coded. These examples are shown in table A.15–A.17. For some variables, the RAs raised issues around consistency in coding. Any additional instructions that we provided in response are also added to this table.

Table A.13: Coding instruction, part 1

var name	meaning	possible codes	coding instructions
claim_yellow_ball	subject says he has received a yellow ball	1	
claim_blue_ball	subject says he has received a blue ball	1	
yellow_no_context	subject says "yellow" without it being clear whether he refers to color of ball received or vote to be cast	1	
blue_no_context	subject says "blue" without it being clear whether he refers to color of ball received or vote to be cast	1	
claim_vote_yellow	a message claiming the sender wil vote for yellow	1	
claim_vote_blue	a message claiming the sender wil vote for blue	1	
sugg_vote_yellow	subject suggests/asks/proposes to vote yellow or claims to vote yellow	1	
sugg_vote_blue	subject suggests/asks/proposes to vote blue or claims to vote blue	1	
sugg_vote_accept	message accepting the suggestion or confirming the suggested/claimed/asked vote	1	
sugg_stm	any message concerning a statement that asks / suggests / proposes to choose a particular statement	VD,D,N,C,VC	statement suggested. If several suggestions/questions are made in 1 message, list them all and separate each claim with ; Use a separate word document to note down the group, period, and session (excel-sheet name) where this happens.
sugg_stm_accept	an acceptance of the suggested / proposed statement, or a message saying that the particular statement will be used	1	
claim_own_stm	any message concerning a statement that deals explicitly and exclusively with what subject himself will state	VD,D,N,C,VC	statement. If several claims are made in 1 message, list them all and separate each claim with ; Use a separate word document to note down the group, period, and session (excel-sheet name) where this happens.
VC_no_context	A message just containing the words "Very Confident"	1	
C_no_context	A message just containing the words "Confident"	1	

Table A.14: Coding instruction, part 2

var name	meaning	possible codes	coding instructions
risk_safe	any message involving reference to risk / gamble or safe choice	1,-1	1 if message suggests to take/consider taking risk; -1 if message suggests not to take/not to consider taking risk
strategy_dec	any message about a strategy to determine what to vote; any message about how to choose what to vote / decide in the future that is not vote trading or experimentation	1	
zero_pay_off_prob	any message about voting blue meaning subjects get zero points, or about voting yellow meaning there being at least a chance of not getting 0 points for sure	1	
vote_trading	in case of a disagreement, existing or anticipated, about what to vote, vote for decision favored by one subject in exchange for voting for decision favored by the other subject in a later round	1	
experiment_dec	any message suggesting experimentation or 'trying something' with votes/decision	1	
link_stm_assess	any message about the relationship between statements and assessments	1	
link_dec_assess	any message about the relationship between decision (Yellow/Blue) and assessments	1	
strategy_stm	any message about a strategy to determine which statement to use; any message about how to choose what statement to use in the future	1	
experiment_stm	any message suggesting to experiment or 'try' something with statements	1	
evaluation	any message concerning previous performance	1	

Table A.15: Coding variables, examples and extra instructions given, part 1.

var name extra instructions	**Examples**
claim_yellow_ball	I have a yellow ball yellow ball NB: "yellow" should be coded as yellow_no_context
claim_blue_ball	similar to claim_yellow_ball
yellow_no_context	yellow
blue_no_context	blue
claim_vote_yellow	I will vote for yellow
claim_vote_blue	similar as claim_vote_yellow
sugg_vote_yellow Typically after an exchange of ball colors	Vote yellow? I think that yellow is the best option yellow? I will vote yellow
sugg_vote_blue	similar to sugg_vote_yellow
sugg_vote_accept Typically after an exchange of ball colors and a suggested decision / vote	OK Yes
sugg_stm NB: code "not very confident" as "confident" NB: code "highly confident" as "very confident"	confident? shall we choose confident?
sugg_stm_accept Typically after a suggested statement or a claim about a statement	OK Yes

Table A.16: Coding variables, examples and extra instructions given, part 2.

var name extra instructions	**Examples**
claims_own_stm NB: code "not very confident" as "confident"	
	I choose confident
NB: code "highly confident" as "very confident"	
risk_safe Typically after subjects have received balls of different colors	Do you like a gamble? Shall we gamble?
Risk / gamble (code 1)	Do you like risk?
NB If after "I have yellow" a subject says "let's take the risk" then this is both a suggested voted and a reference to risk	Just take a chance
If there is only the question "risk or safe" no suggestion is made, risk is discussed	yea lets give it a try (after a proposal to choose yellow with conflicting colors)
If only one option is give, e.g. "let's play safe" then it is a suggestion and a mention of risk/safe	any negative reaction to a suggestion / proposal to play safe / choose Blue
Safe choice (code -1)	
	any negative reaction to a suggestion / proposal to gamble / to choose Yellow Blue gives more in expected terms Blue is the right choice I don't like to gamble / take the risk Its best to pick blue. We will never lose money. choose blue, just to be sure
strategy_dec	
	Let's vote yellow if (reference to color of balls) Let's always do this This is a good strategy / plan Always (if said after a choice for a specific round has been made)

Table A.17: Coding variables, examples and extra instructions given, part 3.

var name extra instructions	**Examples**
zero_pay_off_prob Typically if subjects have received balls of different colors	If we choose blue, we get zero / nothing
	If we choose yellow, we have at least a chance of earning points wel jammer want er moet wel wat binnenkomen jammer dat we geen geld krijgen voor blauw its the only way we have any earnings
vote_trading Typically if subjects have received balls of different colors and they can't agree	
	Let's choose yellow now and choose blue next time Let's choose yellow now and see how it goes; if it doesn't work, we can choose blue next time
experiment_dec	
	let's try this time
link_stm_assess	
	I noticed that we get more points if we say very confident than if we say neutral Shall we say confident to see what assessments we get? Evaluators may think that we are very confident if we tell them we are very confident. They may give us a high percentage Yellow and very confident or just confident. That way we get most out of the evaluators. Okay, i'll think our strategy to vote very confident works. In Round 1 we had a lot of earnings
link_dec_assess	
	I noticed we get more points if we choose yellow we should get more points if we choose blue
strategy_stm	
	Let's always choose very confident Let's choose confident if the colors are different Always (if said after a choice of statement for a specific round has been mentioned)
experiment_stm	
	Shall we choose doubtful to see what happens?
evaluation Typically at the beginning of a round	Did you also get x points?
	They don't like me. That went very well. Using very confident worked! Okay, i'll think our strategy to vote very confident works. In Round 1 we had a lot of earnings what a pity

A.8.3 Coded messages

After both RAs coded every conversation, the coded messages were compared on a line-by-line basis and any disagreement was resolved by the RAs. The result is a consensus version of their coding on a line-by-line basis. This version was aggregated by applying all codes given to messages sent by a specific committee member in a round to that member-round. Table A.18 shows the count of each variable over this dataset. For example, the variable *claim_blue_ball* shows that this type of message was found in 114 member-rounds in the *A-NoStm* treatment. Similarly, committee conversations are all codes applied to both members of the committee in a specific round.

This table forms the basis for Table 5. In Table 5, to determine the percentages in the top part, one should know that 'private signal received' is based on the union of *claim_yellow_ball*, *yellow_no_context*, *claim_blue_ball* and *blue_no_context*; 'vote in that round' on the union of *claim_vote_yellow*, *claim_vote_blue*, *sugg_vote_yellow*, *sugg_vote_blue*, and *sugg_vote_accept*; 'signal received or vote in that round' on the union of 'private signal received' and 'vote in that round' makes '; and 'statement in that round' on the union of the *suggest_vc–suggest_vd* and *claim_vc–claim_vd* variables and *sugg_stm_accept*, *VC_no_context*, and *C_no_context*.

For the lower part of Table 5, we first take the union of both members in a committee for all variables. Then we simply count the number of committees for which *link_stm_assess* (link between statements and assessments), *link_dec_assess* (link between decisions and assessments), *risk_safe* (risk taking), *zero_payoff_prob* (zero-payoff dilemma) are set to 1 for one or more periods. The last variable, united front, is based on the excerpts 4 and 5 in section A.8.1.

In the chat boxes, the signals (balls) received are the most common topic. Members rarely misreport their private signals, whether erroneously or to deceive. Table A.19 shows that members reveal their private signals in almost every member-round. A message in the chat box is coded as *Yellow* if a member specifically claims to have received a yellow ball, or sends a message that just contains the word "yellow" (and similarly for *Blue*).[37] In some conversations, the RAs coded two messages with conflicting interpretations, one indicating a blue and one indicating a yellow ball. This could either be due to a mistake that gets corrected by a member, or an aborted/corrected attempt at deception within the committee, this is coded

[37]That is, it is the union of *claim_yellow_ball* and *yellow_no_context* in Table A.18. Particularly in later rounds, members' conversations become shorter. A typical conversation starts with one member stating "Blue," and the fellow member responding with "me too." Both of these messages would be coded as *Blue*.

Table A.18: Overview of the number of messages coded in a particular category per treatment

	Count		
coded messages	A-NoStm	A-Stm	NoA-Stm
claim_yellow_ball	118	149	232
claim_blue_ball	114	82	268
yellow_no_context	136	109	421
blue_no_context	121	178	364
claim_vote_yellow	15	16	28
claim_vote_blue	9	18	43
sugg_vote_yellow	100	105	168
sugg_vote_blue	102	144	205
sugg_vote_accept	193	204	332
suggest_vc	0	67	10
suggest_c	0	28	10
suggest_n	0	12	2
suggest_d	0	11	1
suggest_vd	0	3	0
sugg_stm_accept	0	100	14
claim_vc	0	13	0
claim_c	0	7	1
claim_n	0	2	0
claim_d	0	2	0
claim_vd	0	0	0
VC_no_context	0	61	2
C_no_context	0	5	3
risk_safe	61	69	127
discus_risk	24	34	68
discus_safe	37	35	59
strategy_dec	18	17	41
zero_payoff_prob	8	8	21
vote_trading	2	1	11
experiment_dec	15	5	24
link_stm_assess	0	33	0
link_dec_assess	12	2	0
strategy_stm	0	23	0
experiment_stm	0	6	1
evaluation	54	118	146
Total	1,983	3,970	6,933

as *Both*. If a member suggested to vote yellow (blue), coded as *sugg_vote_yellow* (*sugg_vote_blue*) in table A.18, but did not reveal his signal before – *i.e.*, is not in *Yellow*, *Blue* or *Both* – we included him in *Suggest yellow* (*Suggest blue*). A message is coded as *Unclear* if a message is sent, but did not enter any of the above categories. Finally, the row *No Discussion* shows in how many member-rounds a member does not use the chatbox at all. Such instances tend to be concentrated around specific members.

Table A.19: Messages sent to fellow committee member as a function of private signal received

	A-NoStm			A-Stm			NoA-Stm		
	Signal			Signal			Signal		
	Blue	Yellow	Total	Blue	Yellow	Total	Blue	Yellow	Total
Yellow	2	204	206	0	217	217	3	562	565
Blue	200	1	201	223	2	225	536	0	536
Both	6	3	9	1	3	4	7	13	20
Suggest yellow	1	0	1	1	18	19	1	1	2
Suggest blue	0	0	0	18	9	27	4	0	4
Unclear	1	1	2	29	22	51	4	0	4
No discussion	0	1	1	46	71	117	3	6	9
Total	210	210	420	318	342	660	558	582	1,140

Notes: *Signal* refers to the signal (ball) that a member received in that round. The rows denote the content of the coded messages in the chat boxes per member-round. *Yellow* (*Blue*) indicates that the member either sends a message claiming to have received a yellow ball (blue ball), or a message that only states "Yellow" ("Blue"). *Both* indicates that a member sends both a message coded as *Yellow* and a message coded as *Blue*; these member-rounds are excluded from the top rows. If a member does not reveal his own signal, so is not included in *Yellow*, *Blue* or *Both*, but does suggest a vote, this is coded as *Suggest yellow* or *Suggest blue*. The row *Unclear* contains all member-rounds in which a message is sent, but the messages fails to meet any of the above. *No discussion* counts all member-rounds where a member sends no message.