TI 2018-070/VII
Tinbergen Institute Discussion Paper

Markets Assessing Decision Makers and Decision Makers Impressing Markets: a Lab Experiment

Revision: May 2019

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

[1] Erasmus University Rotterdam

Markets Assessing Decision Makers and Decision Makers Impressing Markets: a Lab Experiment

Sander Renes and Bauke Visser*

May 3, 2019

Abstract

We experimentally investigate (i) whether markets accurately assess the ability of decision makers when these decision makers benefit from positive assessments and (ii) how decision makers use a costly decision and cheap-talk statements to impress markets. We focus on committees of decision makers to use their conversations as a source of information about their beliefs on the relationship between committee actions and assessments. We find that reputation concerns greatly reduce the amount of useful information markets can rely on. Markets realize this and make assessments less dependent on actual decisions and statements when assessments matter to decision makers. Within treatments, markets use the available information about ability quite efficiently. Reputation concerns make the modal cheap-talk strategy uninformative about ability. In a treatment without statements, committees turn to the decision on the project, a costly signal, to impress. Thus, distorted decisions are more frequent in the absence of the cheap-talk channel.

Keywords: reputation concerns, market assessments, committees, cheap talk, united front, experiment

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, Sacha Kapoor, Debrah Meloso, Lúis Santos-Pinto, Karl Schlag, Otto Swank 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. A previous version was called Committees of Experts in the Lab.

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. These committees operate in environments in which it is often hard to find conclusive evidence as to the 'right' decision, both before and after the decision has been taken. The consequences of many decisions only become clear after years, and the lack of counterfactual information makes forming a judgment about the quality of the decision hazardous.

Short of evidence on the quality of the decision, a natural reaction is to assess the quality of the experts who take the decision and make career-related decisions dependent on assessments. Being considered competent in the eyes of 'the market' or 'the public' is then valuable. The theory of Visser and Swank (2007; VS from now on) predicts that in such a context committees of reputation-concerned experts may take decisions that look good but are bad and that they speak with one voice to convince the market of their expertise. A rational market, however, is not fooled by such behavior. Instead, it anticipates that decision makers have an interest in strategically shaping the information used to evaluate expertise. Nevertheless, committees persist in their behavior to avoid poor assessments.

This interaction between reputation-concerned decision makers and markets evaluating them has been used to explain, e.g., herd behavior, biased forecasts and advice, rash decision making giving way to conservatism, self-censorship in meetings, and various undesired reactions to transparency imposed on committees.[1] Holmström (1999) was the first to analyse this interaction formally. Reputation concerns have also been invoked to explain why the separation of ownership and control can be an efficient form of organization. Fama (1980) argues that managers – decision makers in a firm – take actions with a view to impress the market and that the market for managerial labor is able to discipline managers and induce them to take efficient actions.[2] Others, including Dewatripont et al. (1999a,b), have argued that because of a lack of other financial

[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.

incentives, reputation concerns play an even more important role in the public sector than in the private sector.

Despite its central role in theories concerning decision making and governance, little is known about the equilibrium relationship between decision makers and markets 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 committee takes decisions that look good but are bad, one should know what the right decision is. On the other hand, to measure the quality of the evaluator's assessment one should know the true characteristics of the evaluated decision maker. Neither is easily established by an outsider on the basis of observational data.[3] A second reason may be that the market is modeled as a 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 provide incentives for decision makers to distort their actions in manners not predicted by theory.

To overcome these observability problems, we run a lab 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 able. That is, we replace the mechanistic application of Bayes' rule by human evaluators. The experiment aims to answer two questions. First, how does the presence or absence of reputation concerns affect committee members' behavior – the decisions they take and the cheap talk statements they send to the market – and the quality of the market subjects' assessments? Second, how does the presence or absence of cheap talk as a channel to communicate with the market affect the decision on the project and the assessments of the market? The answers to these questions shed light on the theory of VS and help in assessing the overall functioning of a reputational market – the extent to which market assessments reflect information that is available about the abilities of decision makers.

The set up closely follows the model of VS. In it, committee members receive a private signal about the state, deliberate and vote to either implement or reject a project. All of this happens behind closed doors. Next, the market observes the decision taken – but not the true state – and receives cheap-talk statements from committee members about anything that happened in the meeting. Members care

[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.

both about taking the right, state-dependent decision and about their reputation for expertise. In the model, reputation is defined as the end-of-game probability that a member is of high ability according to the market. In the experiment, it is the role of market subjects to determine this probability. To emulate the reputation concern of committee members, in the experiment part of their payoffs is determined by the probabilities elicited from market subjects.

We find that due to reputation concerns the percentage of committee members whose statements become completely uninformative about their abilities increases from virtually 0% to 30%. 'Very Confident' (in the decision taken) becomes the message that is sent by far the most frequently, whether the members received concurring or conflicting signals. This tendency is in line with the theoretical prediction that cheap-talk communication becomes meaningless with reputation concerns. On average, though, even in the treatment with reputation concerns words speak louder than actions: the statements that members sent contain more information about the underlying ability of committee members than the costly decision on the project. When we shut down the cheap-talk channel in a treatment, members use the decision on the project to manipulate beliefs of the market. This leads to an increase in the frequency of distorted decisions, *i.e.*, decisions that don't maximize expected project payoff.

That members use the decision taken to impress markets is predicted by the model. If members use their private information to maximize the expected project payoff, they should implement the project in case of two positive signals and reject the project in case of conflicting signals or two negative signals. VS call implementation the unconventional decision to stress that it is the a priori less likely decision. Since a high ability member always receives a signal that matches the state, members of a committee that receives two conflicting signals know that at least one of them, and possibly both of them, are of low ability. Thus, if markets believe that committees use their private information to maximize project payoff, they assign a higher reputation to members when the committee implements the project. This creates a tension for reputation-concerned committees. With conflicting signals, the decision that is best from a project-value perspective is to maintain the status quo; the decision that *looks* best is to implement the project. Of course, with rational markets the gains in reputation become lower when committees more actively chase the higher reputation associated with project implementation. Theory predicts that there is an optimal frequency with which committees distort the decision that balances expected losses from implementation in case of con-

flicting signals with gains in reputation stemming from implementation.

In the experiment, the market indeed assigns higher assessments to members of committees that choose the unconventional decision of implementation. Once we control for the cheap talk statements made by committee members, it becomes clear that market assessments react considerably more to statements than to the costly decision. The extent to which they react to statements depends on whether committee members have an interest in obtaining a strong reputation or not. If they have, the market downplays positive statements. Orthogonality tests suggest that market participants make quite efficient use of the information that is available in each treatment, but that they have difficulties in dealing with infrequent statements. We also find that the absence of cheap talk statements causes assessments to be too low by about 7% on average.

Across the treatments, the information that the computer makes available about members' abilities – the pair of private signals about the state of the world – is determined by the same random process. In the last part of the paper we measure how much of that information ends up in the assessments. To do so, we borrow the concepts of entropy and mutual information from information theory. Thus, we measure the amount of information transmitted on a cardinal scale. This allows us to compare the information transmission in absolute terms across the three treatments. We find that, compared with the situation in which members don't care about their assessments, reputation concerns remove 37% of the information about ability that evaluators can potentially glean from observed committee behavior – decisions and statements. The reduction in available information becomes 94% if committees can't use the cheap-talk channel to communicate with the market.

An advantage of studying a committee – rather than a single agent – is that conversations about what decision to take and what statement to send to the market form a natural part of the decision-making process. In the experiment, conversations were computer-mediated, thus creating a source of information on, *e.g.*, decision makers' beliefs concerning the relationship between their actions and assessments. We find that such beliefs determine observed behavior and show that these beliefs can lead to differences with predicted behavior.

In VS, a committee member is either highly able and receives a signal that matches the state or is of low ability and receives a signal that is unrelated to the state. Moreover, a member only knows that he is able with a certain probability. To explain this

relationship to subjects we designed a novel scheme that is based on the traditional urn scheme, see Figure 1. This one picture summarizes all random draws done in a round. The jar at the top contains two balls, a blue and a yellow one. This 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 box one ball is drawn at random, all with equal probabilities. Thus, in the experiment the prior likelihood that a committee member received high quality information, or, equivalently, is of high ability, equals 2/3. In each round in the experiment and for each committee, the computer determines the state of nature, the quality of the information each member receives, and the actual signal that each member receives. The computer does not reveal the color of the ball drawn from the jar nor the letter on the box to any subject. After evaluators have observed committee behavior, they are asked to assess the likelihood that a committee member received high quality information, *i.e.*, a ball drawn from a box labeled H.

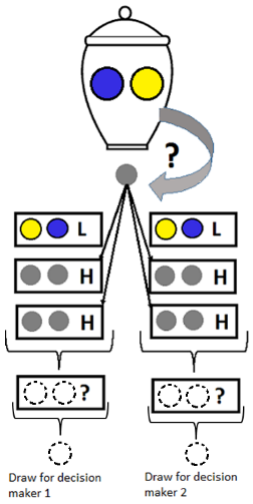

Figure 1: Graphical depiction of relationship between private signals on the one hand and state of nature and ability levels on the other.

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 statements to communicate with an evaluator who gives her assessment after observing the realized state of the world. Moreover, they focus on the behavior of the sender by varying whether the assessments come from computerized evaluators with varying degrees of sophistication or from a human subject. They find that, in line with theory, the more uncertain the state is, the more likely it becomes for the sender to report truthfully. Contrary to theory, they find that assessments react more to the observed accuracy of statements when senders misrepresent their private information than when senders tend to reveal truthfully.

Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2017) study behavior of committees 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. Fehrler and Hughes (2018) find that, in line with theory, transparency hurts decision making as it stifles the free exchange of information in the meeting and makes member unwilling to change their minds. In Mattozzi and Nakaguma (2017), members can differ in two dimensions, ability and bias. They find that whether secrecy or transparency is better depends on the interaction between the two dimensions.[4]

The rest of the paper is organized as follows. We present the theory of VS in the next section and the experimental design in section 3. We discuss the results related to game-theoretic predictions in section 4 and the information-theoretic analysis in section 5. We conclude in section 6. The Supplementary appendix contains additional information about the experiment, various robustness checks and the instructions we used.

[4]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. As we don't study effort, their findings are not directly related to our paper.

2 A theory of decision making by reputation-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 $p < 0$. For this reason, VS call the decision to implement the project the *unconventional* decision.

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 $\Pr(a_i = \bar{a}) = \pi = 2/3$, 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, *i.e.*, a signal that reveals the state, $\Pr(s_i = s^g \mid \mu = h, a_i = \bar{a}) = \Pr(s_i = s^b \mid \mu = -h, a_i = \bar{a}) = 1$. If i is of low ability, he receives a low quality signal, *i.e.*, a signal that contains no information about the state, $\Pr(s_i = s^g \mid \mu = h, a_i = \underline{a}) = \Pr(s_i = s^b \mid \mu = -h, a_i = \underline{a}) = 1/2$. 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 about his private signal $m_i \in \{m^g, m^b\}$ to the other member. 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 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 of high ability on the basis of the observed Y and ω, $\hat{\pi}(Y, \omega) = \Pr(a_i = \bar{a} | Y, \omega)$. We call

this belief the market assessment. It is independent of i.[5] The objective function of a member equals $U_i(Y,\mu) = Y \cdot (p+\mu) + \lambda\hat{\pi}(Y,\omega)$, 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 $(s_1, s_2) = (s^g, s^g)$.

By assumption, members trade off project payoff and the market's assessment in the same way. As a result, there is no conflict of interest within the committee. We follow VS and focus on the equilibrium deliberation strategy of truthful revelation of the private signal. Note that conflicting signals 'cancel each other out' in terms of expected project-value, $\mathbb{E}\left[\mu|s^g, s^b\right] = 0$. This follows from the fact that both members are equally likely to have received an informative signal. A key feature of the model is that receiving conflicting signals also means that at least one member is of low ability. Two high ability members would have received the same, correct signal. We now describe three situations that correspond with the three treatments in the experiment.

First, suppose committee members don't care about evaluations, $\lambda = 0$, but can send statements ω to the market. This corresponds with the No-Assessment-Statements treatment (further: *NoA-Stm*), in the experiment. 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. The market assessment after $Y = 1$ is independent of ω, as $Y = 1$ reveals that $(s_1, s_2) = (s^g, s^g)$. We write $\hat{\pi}(1)$ to highlight the independence of ω. The statement strategy used in case of $Y = 0$ may or may not reveal the pair of private signals. However, as conflicting signals lead to $Y = 0$, it must be the case that $\hat{\pi}(1) > \mathbb{E}[\hat{\pi}(0,\omega)]$, where the uncertainty is over ω. That is, the unconventional decision commands the higher assessment.

Second, suppose that $\lambda > 0$ but 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, $\hat{\pi}(1) > \hat{\pi}(0)$, as the evaluator infers from $Y = 1$ that private signals are the same (and positive), whereas $Y = 0$ may mean that committee members

[5] The independence of i follows from the fact that, because of the assumptions made, a member's private signal *in isolation* reveals nothing about his ability; a comparison of the *pair* of signals is informative, and reveals exactly the same about both members.

received conflicting signals.[6] 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 $\lambda > -p/\left[\hat{\pi}\left(1\right) - \hat{\pi}\left(0\right)\right]$. Define $\beta = \Pr(Y = 1|s_1 \neq s_2) \in [0,1]$ and let $\hat{\pi}_\beta\left(Y\right)$ denote the equilibrium assessment if the committee chooses Y and distorts the decision on Y with probability β conditional on conflicting signals. In equilibrium, β satisfies

$$p + \lambda\hat{\pi}_\beta\left(1\right) = \lambda\hat{\pi}_\beta\left(0\right). \tag{1}$$

Thus, if members have received conflicting signals, what a member gains in assessment thanks to implementation offsets the expected loss due to the distorted decision on the project. As $p < 0$, $\hat{\pi}_\beta\left(1\right) > \hat{\pi}_\beta\left(0\right)$ holds in equilibrium. This requires that, conditional on conflicting signals, committees are less likely to choose $Y = 1$ than $Y = 0$, $\beta < 1/2$. The premium that the unconventional decision commands is lower due to the committees' efforts to "play the system." After all, $Y = 1$ may now result from conflicting signals, while $Y = 0$ is less likely to result from conflicting signals. Thus, $\hat{\pi}_\beta\left(1\right) - \hat{\pi}_\beta\left(0\right) < \hat{\pi}\left(1\right) - \mathbb{E}\left[\hat{\pi}\left(0,\omega\right)\right]$. 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. This situation, with a value of λ such that the committee distorts the decision on Y, corresponds with the Assessment-No-Statement treatment (further: *A-NoStm*).

Finally, suppose that $\lambda > 0$ and 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 on the equilibrium path 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 received. VS draw a second conclusion that is plausible but not dictated by game-theoretic logic: members will show a united front and speak with one voice to the market. Game theory does dictate 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

[6] Thus, $\hat{\pi}\left(0\right)$ equals the expected value $\mathbb{E}\left[\hat{\pi}\left(0,\omega\right)\right]$ if ω is visible.

leads to a drop in assessment. As we use the same value of λ as in the *A-NoStm* treatment, equilibrium voting behavior and market assessments are the same as in *A-NoStm* setting. This situation corresponds with the main treatment in the experiment, the Assessment Statement treatment(further: *A-Stm*).

We use the comparison of the *A-Stm* and *NoA-Stm* treatments to establish the effect of reputation concerns on committee behavior and market assessments. The model predictions for this comparison can be summarized as follows.

1. In the *A-Stm* treatment, the market ignores the statements and bases its assessments solely on decision Y. $Y = 1$, the unconventional decision, commands the higher assessment. In the *NoA-Stm* treatment, the assessment after observing $Y = 1$ is independent of the statements; it may or may not depend on these statements after $Y = 0$, and $\hat{\pi}(1) > \mathbb{E}[\hat{\pi}(0, \omega)]$. The gain in reputation from choosing the unconventional decision is smaller in the *A-Stm* treatment than in the *NoA-Stm* treatment, $\hat{\pi}_\beta(1) - \hat{\pi}_\beta(0) < \hat{\pi}(1) - \mathbb{E}[\hat{\pi}(0, \omega)]$.

2. In both treatments, if committee members receive the same signal, they vote in line with these signals; if they receive conflicting signals, then in the *NoA-Stm* treatment, the committee chooses $Y = 0$, while in the *A-Stm* treatment, the decision is distorted, and becomes $Y = 1$ with probability β that satisfies Eq (1).

3. In the *A-Stm* treatment, a committee member uses a statement strategy that is the same for all pairs of signals; in particular, VS argue that they form a united front. In the *NoA-Stm* treatment, the statement strategy is undefined.

We use a comparison of the *A-Stm* and the *A-NoStm* treatments to establish the effect of the presence of the cheap-talk channel on committee behavior and market assessments. The model predictions for this comparison can be summarized as follows.

1′. market assessments are based on the decision Y, and, conditional on Y, they take on the same value for both committee members, with $\hat{\pi}_\beta(1) > \hat{\pi}_\beta(0)$.

2′. if committee members receive the same signal, they vote in line with these signals; if they receive conflicting signals, the decision is distorted, and becomes $Y = 1$ with probability β that satisfies Eq (1).

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 the experiment, theory predicts that in all treatments members share their private information as conflicts of interest 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 market participant.[7] Our matching schedule needs to balance two goals. The first goal is to avoid 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. 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 2. 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 reducing the time members spend greeting each other and developing a common frame of reference for 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 'group 1' and 'group 2' and the actual members behind 'decision maker 1' and 'decision maker 2' for each committee on an evaluator's screen were randomly determined in every round.

[7] In the experiment, committees were called groups, committee members decision makers while a market participant was called an evaluator.

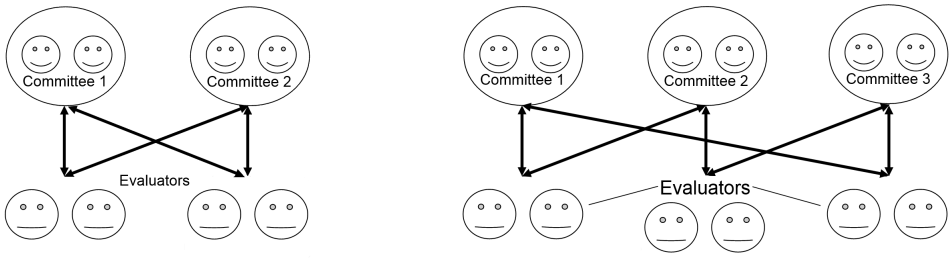

Figure 2: Matching schedules used.

As mentioned in the introduction, we used Figure 1 – both on screen and in the instruction – to explain the relation between the state of nature and the quality of the information, and the actual signals received by committee members. In each round 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 that indicates the quality of the information to anyone.

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 to 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.[8] 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, the equilibrium yields $\beta = 0.33$. 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), see section A.1 for details. 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.

The other two treatments were similar. 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 only. 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, they were prompted to state their degree of confidence in their decision. Next, we ran sessions for evaluators a few days later. Evaluator instructions included the instructions we had given to committee members and, as in the other treatments, explained that it was their role to assess these members. During the experiment, we provided them with the actual decisions and statements

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

of committee members and they were prompted to submit their assessments as in the *A-Stm* treatment. Their incentives were the same as in the other two treatments.

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 ESE-econlab at Erasmus University Rotterdam in the Round September – November 2015. 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 the *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.[9] In all sessions, the first two rounds were practice rounds that could not be selected for payment. Subjects were instructed to use these rounds to get acquainted with the computer environment and the task. In what follows, we drop the first two rounds 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.[10] 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

[9] To check whether the number of rounds matter we do several robustness checks. In Appendix A.6, we show that strategies are qualitatively the same in the first and second half of the experiment in all treatments. We also ran two sessions of the *A-Stm* treatment that lasted for 22 rounds. We find no systematic differences between these sessions and the other *A-Stm* sessions, see Appendix A.7. In the analysis we therefore merge this data with the rest of the *A-Stm* treatment.

[10] 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.

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	375	750	3,000
NoA-Stm	38	42	570	1,140	5,040
Total	110	114	1,155	1,890	9,720

Notes: Number of observations of key variables. CM is the number of committee members, and EV the number of evaluators. Figures for decisions, statements and assessments come from rounds that could be chosen for payment. There are 15 incentivized rounds in the *A-NoStm* treatment. In 3 sessions of the *A-Stm* treatment there are 15 such rounds, while in 2 sessions there are 20. In the *NoA-Stm* treatment we used 30 rather than 15 incentivized rounds to equalize the duration of the treatments.

Table 2: Subject characteristics

	Treatment			
	A-NoStm	A-Stm	NoA-Stm	χ^2
Male	0.5	0.6	0.575	pr=0.475
Year	2.6	3.1	3.01	pr=0.097
Econ background	0.67	0.818	0.86	pr=0.026
Age	20.96	21.3	21.43	pr=0.276
Risk tolerance	5.39	5.4	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.'

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. 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, including the background characteristics in our main regressions does not qualitatively affect the results.

Finally, the chat conversations were independently coded by two research assistants according to a common coding scheme.[11] The two sets of coded conversations were compared and differences resolved by the research assistants.

[11] See section A.10.2 for the coding scheme.

4 Findings

We begin with a discussion of the evaluators' assessments as they form an important part of the payoffs of committee members in two treatments.[12] We then turn to the chat among committee members to see, among others, whether they discuss a relationship between their observable actions and the assessments they obtain. Next, we analyze committee members' behavior.

4.1 Evaluators

We use OLS regressions to determine the weights that evaluators attach to the decision and the statements that committee members make, when determining the assessments. 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	8	0.27	32	0.63
D	88	2.93	368	7.30
N	304	10.13	904	17.94
C	464	15.47	2,344	46.51
VC	2,136	71.20	1,392	27.62
Total	3,000	100	5,040	100

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

As Table 3 shows, the market rarely observes statements that explicitly state doubt, especially if committee members care about assessments. In most of our regressions we therefore cannot control for each statement separately. Instead, we control for the statements 'Very Confident,' 'Confident' and 'Neutral' using dummies, while the statements 'Doubtful' and 'Very Doubtful' are grouped together and used as the low-confidence comparison group. The most extensive model that we estimate is

$$\begin{aligned} A_{ijt} = \alpha &+ \gamma_1 D\left(Y_{it}=1\right) + \gamma_2 D\left(\text{stm}_{it}=VC\right) + \gamma_3 D\left(\text{stm}_{it}=C\right) \\ &+ \gamma_4 D\left(\text{stm}_{it}=N\right) + \gamma_5 D\left(\text{Same Statement}_{it}\right) + FE_t + FE_j + \epsilon_{ijt}, \end{aligned} \tag{2}$$

[12] In the *A-Stm* and *A-NoStm* treatments, the average per-round assessment that a committee member receives equals 68.2 and 60.0, respectively. This should be compared with a decision payoff equal to either 0, 110 or −120.

where A_{ijt} is the assessment by evaluator j of member i in round t, α a constant, $D\,(Y = 1_{it})$ a dummy that is 1 if the committee of which i is a member chooses $Y = 1$ in round t, $D\,(\text{stm}_{it} = VC)$ a dummy that equals 1 if member i in round t uses statement 'Very Confident,' and similarly for stm_{it} equal to C or N, $D\,(\text{Same Statement}_{it})$ a dummy that equals 1 if both members in i's committee choose the same statement in round t, FE_t are round fixed effects and FE_j evaluator fixed effects, and ϵ_{ijt} a zero-mean disturbance term. Note that in every round we redraw a random state, a pair of ability levels and a pair of signals for each committee. These draws create differences between rounds that can influence outcomes. We therefore control for round fixed effects in every regression whenever this is possible. We also control for evaluator fixed effects whenever possible for two reasons. We do this, first, to control for heterogeneity in the response to the scale on which they are required to assess committee members. From survey and experimental research, for instance in the anchoring literature (Furnham and Boo, 2011), we know that individuals can differ in how they respond to such scales. Second, we are not so interested in how average assessments differ across evaluators, but in the way that differences in observed committee behavior cause differences in assessment. We cluster standard errors at the match group level as evaluators in the match group observe the same committees. 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.

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 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 on $Y = 1$. 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 uses 'Neutral' or expresses doubt about the decision, he is likely to be of low ability.[13]

[13]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

Table 4: Assessments as a function of observables

	Assessment				
	A-NoStm	A-Stm	NoA-Stm	A-Stm	NoA-Stm
	(1)	(2)	(3)	(4)	(5)
Y=1	9.602***	4.486*	7.545***	2.428	5.246***
	(1.605)	(2.074)	(1.571)	(1.922)	(1.314)
Very Confident				22.38***	33.38***
				(2.637)	(3.540)
Confident				17.89***	22.95***
				(2.845)	(3.374)
Neutral				6.799***	9.199***
				(1.288)	(2.474)
Same statement				2.242***	1.286
				(0.635)	(0.827)
Constant	54.04***	64.05***	57.56***	46.38***	37.14***
	(1.463)	(3.723)	(2.025)	(3.236)	(2.628)
Observations	1,680	3,000	5,040	3,000	5,040
R^2	0.518	0.296	0.285	0.409	0.583
Cluster level	match	match	match	match	match
Clusters	6	10	21	10	21
Subject FE	✓	✓	✓	✓	✓
Round FE	✓	✓	✓	✓	✓

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* and *Neutral*). *Same Statement* is a dummy set to 1 if this member uses the same cheap-talk statement as his fellow committee member in that round. Robust standard errors are in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Using the same statement as the other group member implies a somewhat higher assessment. In Appendix A.3 we show, using regressions including interaction terms, that evaluators treat decisions and statements as separate sources of information.

A comparison of columns (4) and (5) shows that evaluators make their assessments less responsive to the observed behavior of committee members when the latter are known to care about these assessments. 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.[14]

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.

[14]When pooling the data in columns 4 and 5 and running a regression with a treatment interaction

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 members as an instrument to manage their assessments.[15]

Conclusion regarding predictions 1 and 1′ on assessments. The theory correctly predicts that, across the three treatments, the market gives committee members a higher assessment when they take the unconventional decision. It also correctly predicts that the market reduces its responsiveness to 'positive' information – the unconventional decision, expressions of confidence – when committee members care about their assessments. The prediction that statements are ignored by markets when committee members care about assessments is not borne out by the data. Instead, the market shifts its attention from decisions to statements when statements are made available to committees as an instrument to manage assessments. As we show in the next section, the statements of the majority of members contain information. As a result, the relevant theoretical prediction for evaluators is that they best-reply to this information and integrate it in their assessments. We study whether this is the case in section 4.4.

4.2 Committee members

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

4.2.1 Beliefs and belief formation as obtained from the chat

Table 5 presents key features of the chat.[16] In all treatments, members present the private signals they have received and discuss what to vote. In the *A-Stm* 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. There are also some members who don't use the chat at all. As we show in section A.10, members

on the statement and decision coefficients, an F-test on the significance of the interactions terms rejects the null of no difference between the treatments ($p = 0.0115$). In a test per coefficient, the coefficient on *Very Confident* is significantly different ($p = 0.01$).

[15] A test of no difference of coefficient of $Y = 1$ in the two treatments yields $p < 0.01$.

[16] Table A.30 presents a complete overview of all coded variables.

who chat about their private signals or intended vote virtually always truthfully reveal their signals.

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. The payoff-irrelevance of the statements in the *NoA-Stm* treatment makes that they are hardly a topic for conversation. In the *A-Stm* treatment on the other hand, most committee members write about statements, but the frequency goes down over time.

The lower part of the table shows what members believe to be driving the assessments they receive. 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 at some point during the experiment relate the assessments they receive to the statements they make. In this treatment, only 2 committees ever relate the assessments to the decisions. As a result, members more often discuss what statements to choose in a bid to affect their assessments than what decisions to take. This 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 in the *A-Stm* treatment that discuss the relationship between their statements and assessments sheds light on how committees come to form these beliefs. 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.[17]

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 the market.

[17]See excerpts 1 and 2 in Appendix A.10.1 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	77.1	99.6
- vote in that round	76.7	51.5	52.6
- signal received or vote in that round	100	83.5	99.8
- statement in that round	–	31.2	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	3	–
- united front	–	2	–
- risk taking	12	17	19
- zero-payoff dilemma	5	7	12

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. The statistics in the top part of the table are based on the incentivized rounds, whereas the numbers in the lower part also includes the first two practice rounds. We include the practice rounds as part of subjects' understanding of the game may take place in these rounds.

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.[18] By this test, we find little support for the corresponding part of prediction 3: only two committee pass it, see excerpts 4 and 5 in Appendix A.10.1.

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.

[18] 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 statement.

Discussions about what drives assessments and the zero-payoff dilemma have consequences for committee behavior as we discuss in the next two sections.

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 3 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 25.9% of the time in the *A-Stm* treatment, a percentage that is indeed significantly lower than the 37.1% in the *A-NoStm* treatment.[19] 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. Committees that don't discuss this dilemma or have not yet discussed it choose $Y = 1$ in 24.4% of the rounds with conflicting signals. Once committees do discuss the dilemma, this percentage jumps to 60%.[20] In the *A-Stm* treatment, the zero-payoff dilemma is practically not discussed. Moreover, assessments are associated with statements, hardly with decisions. It would thus be consistent with the discussions to observe that more committees distort the decision on Y in the *NoA-Stm* treatment than in the *A-Stm* treatment. Figure 3 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.[21]

In sum, the beliefs, as obtained from the chat, about the determinants of assessments have consequences for the decisions taken. This drives a wedge between the predicted frequency of distortions and our findings.

Conclusion regarding predictions 2 and 2′ on the relationship between signals and committee decisions. In line with theory, and across the three treatments, if members receive the same signal, committees take the decision that matches their signals. In case of conflicting signals, the predicted frequency of distorted decisions

[19] $p = 0.03$ in a two-sided test of proportions.
[20] $p < 0.001$ in a two-sided t-test with unequal variances.
[21] $p = 0.05$ in a two-sided test of proportions.

does not correspond with the findings. A comparison of the *A-Stm* and *NoA-Stm* treatments shows that he presence of reputation concerns leads to fewer distortions if committees can use cheap talk statements to communicate with the market, rather than more. A comparison of the *A-Stm* and *A-NoStm* treatments shows that the possibility of using cheap talk statements reduces the frequency of distortions, rather than leaving it unaffected. We showed that these patterns are consistent with members' beliefs about what determines their assessments and presence or absence of reference to the zero-payoff dilemma. There is no zero-payoff dilemma in the theory of VS. In line with the prediction, committees in the *NoA-Stm* treatment that don't discuss this dilemma are less inclined to distort the decision than committees in the two treatments with reputation concerns.

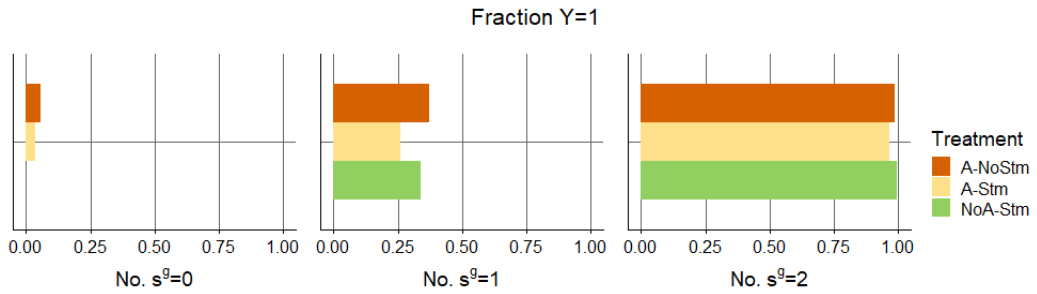

Figure 3: Relationship between number of positive signals and committee decision

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

4.2.3 Consequences for statements

Prediction 3 says that a concern with assessments causes cheap talk statements to be uninformative about members' abilities. We find that, in line with this prediction, a concern with assessments greatly reduces the variation in the statements used and their informativeness. Moreover, over time, a growing number of members reveals less about their private signals – and thus about their abilities – through their statements. There is also quite some heterogeneity among members: many members do reveal information about their ability levels through their statements.

Figure 4 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 (*NoA-Stm*), frequency distri-

butions are roughly bell-shaped. When they have received the same signals, members' modal statement is 'Confident.' They use this statement 62,7% 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.

We know from the chat that in the *A-Stm* treatment members relate the assessments they obtain to the statements they make, hardly to the decision they take. This 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 81.4% of the time if their signals agreed, and more than 44% of the time if they have received conflicting signals. Statements expressing doubt become extremely rare.

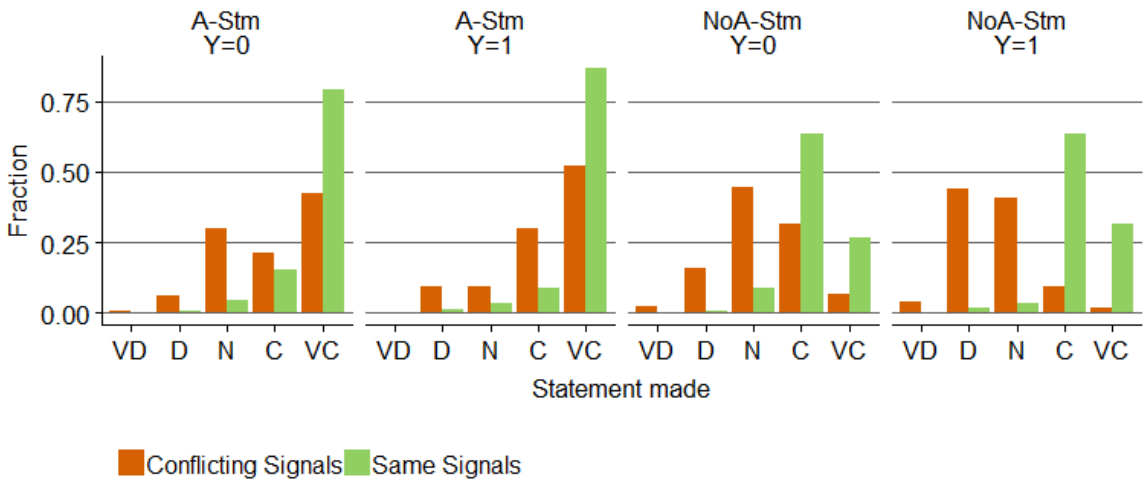

Figure 4: Observed statement strategies

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.

Further evidence that the belief that statements affect assessments actually changes members' behavior can be obtained from observing how the statements that a committee member uses change when this belief is articulated by him or his fellow committee member. This is illustrated in Figure 5. To make this figure, we score every statement

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 below the line 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 above the line 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 members increase their stated confidence when they relate assessments to statements.

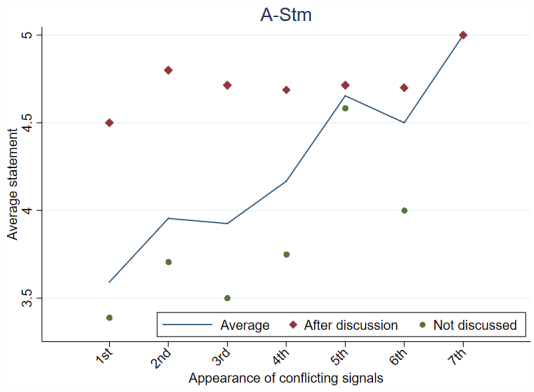

Figure 5: Discussing the link between statements and assessments increases the stated level of confidence

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.

Members form their beliefs over time and beliefs shape behavior. This has consequences for the distribution of statements in the *A-Stm* treatment. In the first five incentivized 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 rounds 12 till 17 the frequency with which members

report 'Confident' or 'Very Confident' has gone up to 79% in case of conflicting signals (difference in two-sided test of proportions, $p = 0.003$), with 'Very Confident' causing the bulk of the increase. Conditional on having received the same signals, the usage of these two statements remains constant, at 93%. The only change that happens 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.[22]

A concern with assessments makes that members express considerably higher levels of confidence and this tendency becomes stronger over time. Some heterogeneity among committee members remains. To illustrate this, 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 4, with 0 meaning that, on average, a member uses the same statement in both types of rounds. The score 4 can only result from a member having averages equal to 1 and 5 for the two types of rounds. That is, the lower is the score, the less distinct the statements in the two types of rounds are.

Figure 6 shows the distribution of members' transparency. The vertical lines indicate the cut-offs between terciles. The modal transparency score in the *A-Stm* treatment, attained by more than 30% of the CM, is zero. As a result, 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* treatment is quite different. In fact, the modal difference is 1 full point and many members are even more transparent. Compared with the *NoA-Stm* treatment, a concern with assessments reduces the average transparency by 0.41 on a 5-point scale.[23]

Conclusion regarding prediction 3 on statements. In line with the prediction, the modal degree of cheap-talk transparency is zero when members care about assessments. Moreover, over time, a growing number of members reveals less about their private signals through their statements. This shows that reputation concerns indeed make statements considerably more uniform across signal pairs. Nevertheless, many

[22] 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 rounds, while these percentages equal 90% in the first round and 94% in the last, conditional on members having received the same signals.

[23] This reduction is statistically significant, with $p < 0.001$ in a two-sided t-test.

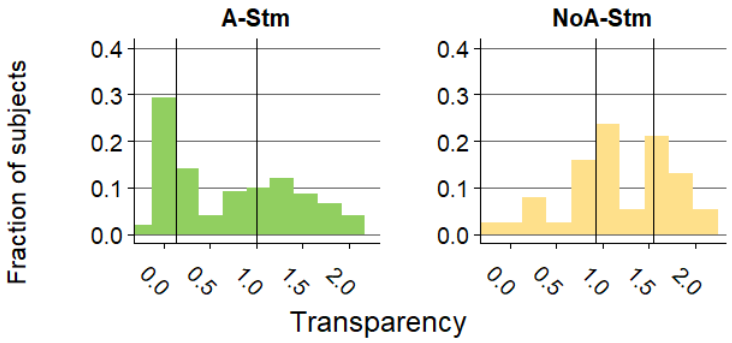

Figure 6: Members' cheap-talk transparency

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.

members do reveal information about their ability levels through their statements. In section 5 we calculate the mutual information of ability given statements to have a formal measure of the amount of information that the statements contain.

4.3 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. There is ample evidence that this is indeed happening. In the *A-Stm* treatment, in which members related the assessments they receive to the statements they choose, statements are markedly different from those used when members are unconcerned with assessments. When we shut down the statement channel, as in the *A-NoStm* treatment, members' attention shifts to decisions and the incidence of distortionary decisions goes up. At the same time, there is considerable heterogeneity in members' behavior. This naturally raises the question what type of behavior was most suited given the way the market reacted.

Members can exhibit heterogeneity in what statement they use and in what they decide as a committee. We reported on the former form of heterogeneity, using transparency, in section 4.2.3. Committees 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 his vote*.[24] In what follows, we distinguish between members who do and who don't distort their vote. Table 6 shows the joint distribution of both forms of heterogeneity. How does a concern with assessments affect this joint distribution? In the *A-Stm* treatment there seems to be a division between committee members who act strategically and have both a low level of transparency and distort their votes, and those who 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 6: Cheap-talk transparency and distorted votes

	Cheap-talk transparency tercile					
	A-Stm			NoA-Stm		
Inclination to distort	bottom	middle	top	bottom	middle	top
no	2	7	7	2	2	1
yes	13	7	8	11	10	12
Total	15	14	15	13	12	13

Notes: Joint distribution of members' cheap-talk transparency and inclination to distort their votes. 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 their vote is defined as the fraction of rounds with conflicting signals in which the member votes v^1. If this fraction equals zero, then inclination to distort equals no; if fraction is positive, then inclination to distort equals yes.

Given actual project payoffs and market's assessments, what type of behavior as defined by a member's transparency and his inclination to distort gives a committee member the highest total payoff? Table 7 shows how project payoffs and average assessments vary with member's choices on both dimensions (in the *A-Stm* treatment) or only with their inclination to distort their vote (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 their vote in this bottom tercile of transparency 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

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

would raise suspicion with evaluators and possibly harm their average evaluations, see excerpts 6–8 in section A.10.1. The data shows that such a concern was not only unwarranted but even harmful if the committee member, as a result of this concern, used a statement strategy with positive transparency.

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 rounds, 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, 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 7: Average payoffs received by committee members

	Transparency terciles A-Stm			A-NoStm
	Bottom	Middle	Top	
Inclination to distort, no				
Project payoff	31.4	36.6	37.6	35.4
Reputation	69.3	65.4	67.4	58.4
Total	100.7	102.0	105.0	93.8
Inclination to distort, yes				
Project payoff	21.0	27.7	24.5	25.8
Reputation	70.8	68.8	66.7	60.8
Total	91.8	96.4	91.2	86.6

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

4.4 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.

This section uses an orthogonality test to shed light on any systematic mistakes made by evaluators in transforming the information they observe into their assessments. It is based on Keane and Runkle (1990, 1998)'s study of the rationality of individual forecasts of prices and profits. Following these papers, 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.

To test for unbiasedness, we use the two-sided t-test in Table 8. The t-tests show that in the treatment without statements, *A-NoStm*, the average assessment is too low by about 4 percentage points. This corresponds with roughly 7 % of the average assessment – significant, but not extremely large from an economic point of view. In the treatments with cheap-talk statements, differences are considerably smaller. Some bias is present in the *NoA-Stm* treatment, but the difference is insignificant in case of *A-Stm*.

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	<0.001	1,680
A-Stm	68.23	16.21	67.73	46.76	0.50	0.572	3,000
NoA-Stm	65.29	17.61	66.90	47.06	-1.61	0.016	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.

The efficiency test of Keane and Runkle (1990, 1998), when applied to our setting, amounts to testing whether any behavior of committee members that is observed by evaluators has predictive power for members' true ability over and above evaluators' assessments. One cannot simply use an OLS regression of actual ability on assessments

and observable committee behavior because the dependent variable is binary, while assessments are continuous. However, if information is used efficiently, there should be no systematic link between observed committee behavior and the expected value of the *difference* between the assessment and ability.[25] We therefore 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 is of low ability in round t. This variable captures ability, and is defined on the same percentage-points scale as the assessments of the evaluators so that they can be directly compared. We then define the variable *Mistake*, as $\Delta_{ijt} = h_{it} - A_{ijt}$, and see how it relates to observable signals in an orthogonality test of the form:

$$\Delta_{ijt} = \alpha_0 + \alpha_1 X_{ijt} + \epsilon_{ijt}, \tag{3}$$

where X_{ijt} captures all behavior of the committee of which i is part that j observes in round t. Efficiency requires that $\alpha_0 = 0$, so that the average difference is 0, and all $\alpha_1 = 0$ so that we find no systematic relation between observable signals and mistakes in assessments in this regression.

In our experiment there is a strong correlation between the information that is available about the two members of a given committee in every round, as they have taken the same decision Y 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 rounds or within matching groups. They show that one obtains a consistent estimate of the variance of the coefficients by clustering the standard errors.[26] For our experiment this implies a cross-sectional cluster on the level of the matching group and a temporal cluster on the round. As we have only a limited number of clusters, we bootstrap these clusters using the wild bootstrap procedure of Cameron et al. (2008). This bootstrap procedure estimates error terms within clusters, so that we need sufficient variation of the X_{ijt} variariables in each cluster. All of our explanatory variables are dummies that have limited variance. Furthermore, as we noticed before, low confidence statements become less frequent in the final rounds of the experiments. We are forced to drop the *Neutral* category and merge it with the low confidence statements. We also drop the *Same Statement*

[25] If one views the assessment as a prediction of ability, efficiency means that, conditional on observed committee behavior, the expected value of the prediction error is zero.

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

variable because the lack of variance in some clusters makes it impossible to bootstrap the standard errors. Full regressions with all variables with non-bootstrapped standard errors are shown in Table A.23 in the Appendix. They yield the same results qualitatively, but with smaller estimated standard errors.

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

	Mistake					
	Two-way clustering			No clustering		
Y=1	-1.375	-3.463	2.221	-1.375	-3.463*	2.221
	(3.894)	(4.385)	(4.144)	(2.421)	(1.805)	(1.356)
Very Confident		9.110	2.745		9.110***	2.745
		(7.267)	(5.264)		(2.652)	(1.844)
Confident		8.714	8.752*		8.714***	8.752***
		(10.05)	(4.897)		(3.290)	(1.639)
Constant	4.893*	-6.882	-4.203	4.893***	-6.882***	-4.203***
	(2.873)	(5.767)	(4.209)	(1.662)	(2.449)	(1.381)
Observations	1,680	3,000	5,040	1,680	3,000	5,040
R^2	0.000	0.005	0.007	0.000	0.005	0.007
Cluster level	Match & Round	Match & Round	Match & Round	-	-	-
Subject FE	-	-	-	-	-	-
Round FE	-	-	-	-	-	-

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*). Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Columns (1)–(3) of Table 9 presents the results with bootstrapped, two-way clustered errors. In general, there are few signs of significant and systematic errors. The coefficient of $Y = 1$ is not significant in any treatment. This means that there is no sign that the market systemically mistreats any information about members' abilities that can be obtained from the decisions taken by the committees. The constant in the test for the *A-NoStm* treatment is bordering on the 10% significance level, which suggests that evaluators, on average, give biased assessments. This is consistent with what we have seen above. There are no signs of systematic mistakes in the main treatment, *A-Stm*. In the *NoA-Stm* treatment, the coefficient of *Confident* is statistically significant, but its sign is opposite to the constant. To interpret this pattern, recall that *Confident* is the modal statement in this treatment. If we test whether the sum of the constant

and that coefficient is different from 0 we get a non-significant result. This coefficient therefore seems to indicate that the relatively *uncommon* statements are assessed wrongly. Overall, 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 perform the same orthogonality test, with the same clusters, but now include 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.23 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 part, the lack of significance of the coefficients of observable variables stems from the fact that the two-way clustering creates large standard errors. We therefore reproduce these regressions without the clusters in columns (4)-(6). These regressions heavily overestimate the independent information in every observation. The coefficient on $Y = 1$ is not significant in any treatment. Thus, even without the clustered standard errors there is no sign of systemic mistreatment of the information contained in the committees' decisions. Clearly, the bias in the *A-NoStm* treatment remains. In columns (5) and (6), three out of four of the coefficients on *Very Confident* and *Confident* are significant, but have a sign opposite to the constant in both the *A-Stm* and *NoA-Stm* treatment. Note that the statements 'Confident' and 'Very Confident' account for 73 % of all statements observed by the market in the *NoA-Stm* treatment and for 85 % of all statements in the *A-Stm* treatment. In the *A-Stm* treatment, an F-test on the restriction that the coefficient on *Very Confident* plus the constant equals zero in the regression without clusters, column (5), is not rejected($p = 0.4298$). The same test for the second most likely statement, 'Confident,' gives a p-value of 0.96. This indicates that, on average, the market assesses (close to) correctly 85 % of the statements. Systematic mistakes seem to be confined to infrequent statements. In the *NoA-Stm* treatment and the regression without clusters, column (6), an F-test on the restriction that the sum of the coefficient of *Very Confident* and the constant equals zero rejects the null-hypothesis ($p = 0.0001$). The same test for the most likely statement, 'Confident,' gives a p-value of 0.3208. So again, the most used statement is

assessed correctly on average.

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 for uncommon pieces of information and as a bias, *i.e.*, at the *average* level of the assessments.

5 Entropy

The orthogonality tests presented in section 4.4 examine the extent to which the market makes rational use of observed committee behavior within a given treatment. These tests don't control for the amount of information that evaluators actually observe in a treatment. As a result, they cannot be used to compare the degree to which the information that the computer makes available in each round about members' ability – the absence or presence of conflicting signals – is transformed by members' behavior and how much of that information eventually finds its way into the assessments. However, such a comparison is relevant for our study as it sheds light on the way in which the presence or absence of reputation concerns and of cheap talk determines the amount of information that is available for the market to base its assessments on. 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. To do so, we measure the available amount of information in each treatment on the cardinal entropy scale.[27] 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 $-\log_2 p_i$ and is measured in bits. Thus, the less likely an outcome is, the higher its information.

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

The entropy of variable X is defined as the expected information of X,

$$H(X) = -\sum_i p_i \log_2 p_i. \tag{4}$$

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 entropy that remains about a variable X after observing another variable Z. This is measured by the conditional entropy,

$$H(X|Z) = -\sum_j p_j H(X|z_j). \tag{5}$$

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

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

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 10 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).[28] 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 *Stm4* variable captures the statements. As before, we bin the lower two statements. *Info Set* is a variable that combines *Y=1* and *Stm4*, creating a variable with 8 possible values. *Info Set2* is a variable that combines Y and the *Stm4* variable of both committee members in a committee, with 32 possible values. *Asessement* is the assessment of an evaluator.

Table 10: 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	Stm4	Assessment
A-NoStm	0.9407	0.0706	0.0050	0.0050	0.0050	-	0.0198
A-Stm	0.9075	0.0941	0.0510	0.0245	0.0002	0.0241	0.0099
NoA-Stm	0.9160	0.0938	0.0815	0.0606	0.0109	0.0495	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 $Y = 1$. *Stm4* codes the four levels of statements we use {'Low,' 'Neutral,' 'Confident,' 'Very Confident'}, where 'Low' combines 'Doubtful' and 'Very Doubtful.' *Info Set* combines the information in Y and *Stm4* in a single categorical variable with 2×4 categories. *Info Set2* combines the information in Y and the *Stm4* variables of both committee members in a single categorical variable with $2 \times 4 \times 4$ 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.24 in the Appendix reports the bootstrapped standard errors of these estimates.

In all treatments, members' ability as determined by the computer has a similar

[28] 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.24 in the Appendix, we report the bootstrapped standard errors of these estimates. Calculations were made using the 'infotheo' package in *R* of Meyer (2014).

level of entropy, see column (1). The pair of private signals that members receive significantly reduces the entropy, see column (2). Still, considerable uncertainty 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). A concern with assessments considerably reduces the amount of information about ability that committee members reveal through their behavior, and this reduction is larger in the absence of cheap talk statements. Indeed, the mutual information of ability given *Info Set2* or *Info Set* is largest in the absence of a concern with assessments, 0.0815 and 0.0606 bits, respectively. It is only about half that size in the *A-Stm* treatment, 0.0553 and 0.0311 bits, respectively, and more than 10 times smaller in the *A-NoStm* treatment, 0.0050 bits.

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 is weak since only a limited amount of information about ability is available to the committee members themselves (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.

To sum up, when committee members care about their assessments they obfuscate the signals they receive, considerably complicating the market's job of assessing them. 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 a concern for assessments is 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 costly actions. This justifies the dependence of assessments on these statements as we found in section 4.1.

6 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 of the world when assessing committee members. The controlled lab environment allows us to discern whether committee members take a decision that looks good but is bad and to establish the determinants and measure the quality of the assessments of evaluators. We use the chat within the committees to shed light on what the decision makers perceive to be the incentives in the reputational market, *e.g.* the beliefs about the relationship between assessments and actions taken or statements made. The chat also allows us to see whether the statements sent to evaluators are coordinated or not.

We find ample evidence that a concern with assessments changes the behavior of members. When they have reputational concerns, members express greater confidence and the modal statement strategy becomes pure, *i.e.* meaningless, cheap talk. The majority of members, though, still condition their statements on their private signals, albeit to a lesser degree than in the absence of reputation concerns. This means that even with such concerns, evaluators can glean information about ability from the statements they observe.

We also find that the market takes these behavioral reactions into account when assessing committee members. As predicted by theory, in all treatments evaluators assess the ability of committee members higher when committees take the unconventional decision to implement rather than 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. Thus, although reputation-concerned committee members interfere to affect the inferences drawn by the market, the market still manages to use the available information quite well.

Even though our orthogonality tests show that the market uses available information quite well, they do not show how much information is available in the first place. We measured the available information using the cardinal entropy scale. This information-theoretic analysis shows that reputation concerns reduce the amount of information about ability embedded in the committee's observable actions by about one third. As a result, evaluators must do with less information exactly when assessments matter. This analysis further shows that in this experiment, words speak much louder than costly actions. The entropy measures thus show that the reputation market has

much more information available, and thus can work more effectively, if the decision makers can also use cheap talk.

The 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. In the main treatment a substantial number of members scores low on transparency and, with some frequency, takes the decision that looks good but is bad. An equally large group, however, maintains a higher level of transparency and avoids distortions. The heterogeneity of strategies used by our subjects implies that careful selection can alleviate the tension between the need for reputation incentives and the effectiveness of the reputation market. It is common to find heterogeneity in types of behavior observed in experiments. We are not aware, however, of any theory on committee behavior that makes this part of its assumptions.

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.[29] However, claiming to be 'very confident' in the decision even though one's committee received conflicting signals is different from lying about one's private signal. We still find that many subjects are not willing to use the statements completely strategically and, thus, the statements contain information. 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. Theory predicts that only costly decisions are effective in influencing assessments. We find that cheap talk is a substitute for the costly decision as a means to influence assessments. The result is less distortions in decisions when cheap talk is available. This suggests that if real-world decision makers care about their assessments, one should require their decisions to be accompanied by statements.

In our experiment several random draws interact. These interactions could be complicated to subjects that are unfamiliar with the environment and thrust into their experimental roles for the short duration of the experiment. Our urn scheme provides a simple way of communicating these draws and their interactions. The three treatments also suggests a strategy for acquainting subjects with a full game. Instead of using

[29] 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.

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 simplified 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

Furnham, Adrian, and Hua Chu Boo (2011) 'A literature review of the anchoring effect.' *The journal of socio-economics* 40(1), 35–42

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