Strategic Expert Committees and the Markets That Assess Them

A laboratory experiment

Renes, Sander; Visser, Bauke

DOI
10.1287/mnsc.2021.03007

Publication date
2024

Document Version
Final published version

Published in
Management Science

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 71, No. 7, July 2025, pp. 6227–6247
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Strategic Expert Committees and the Markets That Assess Them: A Laboratory Experiment

Sander Renes,[a,*] **Bauke Visser**[b]

[a] Faculty of Technology, Policy and Management, Delft University of Technology, 2628 BX Delft, Netherlands; [b] Erasmus University Rotterdam and Tinbergen Institute, 3062 PA Rotterdam, Netherlands

*Corresponding author

Contact: s.renes-1@tudelft.nl, https://orcid.org/0000-0003-2495-9219 (SR); bvisser@ese.eur.nl, https://orcid.org/0000-0002-3212-3684 (BV)

Received: September 6, 2021
Revised: June 30, 2023; March 1, 2024; May 31, 2024
Accepted: July 13, 2024
Published Online in Articles in Advance: October 28, 2024

https://doi.org/10.1287/mnsc.2021.03007

Abstract. Committees of experts are widely used to make decisions. We experimentally investigate the relationship between decision making in committees and the assessment of the ability of committee members by evaluators, comparing observed behavior with theoretical predictions. Treatments vary in whether members care only about a state-dependent project payoff or also about assessments and whether evaluators can base their assessments only on the decision the committee makes or also on cheap-talk statements made by committee members on their confidence in the committee decision. Evidence for the equilibrium predictions is mixed; for example, contrary to theory, committees with a concern for their assessment do not distort their decisions more than committees without, whereas in line with theory, evaluators give higher assessments to committees that take the risky decision rather than the riskless. We analyze whether evaluators rationally base their assessments on observed behavior of committees using an orthogonality test. In treatments with cheap-talk statements, assessments are quite rational; in treatments without, they are too low on average. We investigate whether committees best reply to expected project payoffs and, in treatments in which members' payoffs also depend on assessments, to predicted assessments conditional on observed committee behavior. In all treatments, committees respond to the possibilities to raise expected payoffs but do not use them as much as predicted by theory. We conclude by showing that the amount of information about committee members' abilities that ends up in assessments is considerably larger when evaluators observe committees' decisions and statements rather than only decisions.

History: Accepted by Yan Chen, behavioral economics and decision analysis.
Funding: We gratefully acknowledge financial support from the Dutch Research Council (NWO) under the grant 400-09-338; and Erasmus University Rotterdam [Grant CSTO 2014–54].
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2021.03007.

Keywords: reputation concerns • market assessments • committees • cheap talk • experiment

1. Introduction

Decision-making committees—such as corporate boards, monetary policy committees, or health consensus panels—are frequently used to bring together experts on a specific matter. These expert committees operate in contexts in which it is 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 even then, the frequent lack of counterfactual information makes judging the quality of the decision—and of the decision makers—difficult. In the absence of conclusive evidence about their qualities, assessments of the decision makers for, say, retention or promotion, can only be based on the decisions they took and the statements they made to explain their choices.

Since Fama (1980), the relationship between decision makers and reputation markets has played a central role in theories about governance. Key elements of these theories, such as the ability of decision makers, the correctness of the decision, and the informational content of decision makers' statements, are difficult to establish using observational data, making it hard to empirically test these theories.[1]

To overcome these observability problems, we run a laboratory experiment. In the experiment, half of the subjects form two-member committees that make a binary decision under uncertainty, whereas the other subjects evaluate whether committee members are able.

The experiment aims to answer two sets of questions. First, how does the presence or absence of reputation concerns affect committee members' behavior and the

quality of the assessments of the evaluators? Second, how does the presence or absence of cheap-talk statements by committee members as a source of information on which evaluators can base their assessments affect the decisions of the committee and the assessments of the evaluators?

To answer these questions, we compare behavior and assessments in a 2×2 treatment design with treatments varying in whether members care only about a state-dependent project payoff or also about assessments and whether evaluators can base their assessments only on the decision the committee makes or also on cheap-talk statements made by committee members on their confidence in the decision made. As the experiment closely follows the model of Visser and Swank (2007) (VS), we can compare theoretical predictions and experimental findings.

In all treatments, each member receives a signal about the state. A member's signal in isolation does not reveal information about the member's ability, but the pair of private signals does: conflicting signals are a sure sign that at least one member, perhaps both, is of low ability. 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 evaluators to assess this probability. To emulate the reputation concerns of committee members in the experiment, part of their payoff is determined by the assessments elicited from evaluators.

An important prediction of the VS model that we check in every treatment is that, in equilibrium, one decision (project implementation) commands a higher reputation than the other (rejection). As a result, committee members face a trade-off when they have received conflicting signals about the state: from a project-value perspective, rejection is best; from a reputation perspective, implementation is. This leads to an important predicted treatment effect: members who are unconcerned about their reputations reject the project; members who are concerned about their reputations are willing to give up project value for a stronger reputation. In the latter case, rational markets see through this incentive to take the decision that looks good and rationally reduce the gain in reputation from project implementation. Nevertheless, in equilibrium, even when reputation-concerned members distort the decision in the case of conflicting signals by implementing the project with some positive probability, project implementation commands a higher reputation than maintaining the status quo.

When we compare treatments with and without cheap-talk statements, the model predicts the same difference in reputation between implementation and rejection. When committee members care about their reputations, statements are uninformative about members' abilities: members use a pooling strategy when it comes to statements, using the same statement strategy whether they received conflicting signals or the same signals. The reason is that, had one statement rationally induced the market to reward a committee member with a higher reputation, then members would have opportunistically exploited this possibility of raising their reputations irrespective of what they know themselves about their abilities. As a result, the model predicts that evaluators ignore the statements they receive and instead base their assessments only on the decision made by the committee—a costly signal—exactly as in the treatment without statements. Thus, the difference in reputation between implementation and rejection is the same whether evaluators can or cannot base the assessments on cheap-talk statements.

When members do not care about their reputations, there are no payoff consequences of the statements, and the model of VS, therefore, does not predict how informative cheap-talk statements are in this setting. Statements cannot, however, become less informative than completely uninformative, so intuitively, the expectation is that there should be more information in the statements when members do not care about their reputations than when they do. This intuition is also suggested by earlier experiments that find, even when there are reasons to hide information, participants tend to communicate honestly at least to some degree, leading to what has become known as overcommunication (Dickhaut et al. 1995, Cai and Wang 2006, Goeree and Yariv 2011, Fehrler and Hughes 2018, Meloso et al. 2023).

We find mixed evidence for the predictions of VS. For example, on average, assessments are indeed higher after implementation than after rejection in every treatment, and this difference is indeed larger in treatments without reputation concerns than with. However, the predicted difference in frequency of distorted decisions—implementation in the case of conflicting signals—is not found in the data: committees who do care about their reputations distort the decision as often as committees who do not care. Moreover, in the treatment in which payoffs of committee members increase in the assessments of evaluators, evaluators' assessments and committees' behavior differ from equilibrium predictions: evaluators raise their assessments when committee members express more confidence in the decision made, whereas about half the committee members use statement strategies that depend on the private signals they have received and are, therefore, at least somewhat informative about members' abilities.

To understand these deviations, we investigate whether committees best reply to expected project payoffs and, in treatments in which members' payoffs also depend on assessments, to observed assessments. We find that, in all treatments, committees respond to the possibilities to raise expected payoffs but do not use them as much as predicted by theory. We also find

indications of behavioral effects in the laboratory that cause more distortions in the treatments without reputation concerns but not in the treatments with. To understand whether evaluators rationally base their assessments on observed behavior of committees, we use the orthogonality test proposed by Keane and Runkle (1990, 1998). We find that, in treatments with cheap-talk statements, assessments are quite rational. For example, evaluators correctly downplay both the decision on the project and positive statements when forming assessments in treatments with reputation-concerned committee members. However, in treatments without cheap-talk statements, assessments are too low on average.

At the start of every round in every treatment, the computer sends both committee members a signal about the state of the world. This pair of signals is the only information available from which any subject in the experiment—committee members and evaluators—can infer something about the ability levels of committee members. Clearly, committee members who share their private signals observe the pair; evaluators do not directly observe this pair as they only observe committee decisions and, depending on the treatment, cheap-talk statements. In the last part of the paper, we measure how much of the available information about ability 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 treatments. We find that the amount of information about committee members' abilities that ends up in assessments is considerably larger when evaluators observe committees' decisions and statements rather than only decisions.

In this experiment, on average, words speak louder than actions and they are more consequential for assessments. For the treatment with reputation concerns, this is the opposite of what theory predicts.

In theory, the reputation market acts as a machine that dutifully applies Bayes' rule to the equilibrium behavior of decision makers to determine the resulting reputation (Holmström 1999). Human evaluators, however, may struggle to interpret the actions of decision makers, especially when these decision makers care about the assessments. Furthermore, subjects may treat strategic uncertainty stemming from the actions of others differently than objective, nonintentional risks because of other random effects (Butler and Miller 2018; Chierchia et al. 2018; Li et al. 2018, 2019, 2020). Any difficulty in applying Bayes' rule can cause evaluators' assessments to differ from theoretical predictions and, thus, cause experts' incentives to differ from theoretical predictions as well. Our informational efficiency test suggests that evaluators make rational use of the information available to them in the treatments in which they also receive cheap-talk statements; in the treatments in which they do not, assessments are biased.

The interaction between reputation-concerned decision makers and the markets evaluating them is first formally analyzed by Holmström (1999). His analysis was inspired by Fama (1980), who argues that managers take actions with a view to impress the managerial labor market and that this market is able to discipline these managers. Dewatripont et al. (1999a, b) argue that, because of a lack of other financial incentives, reputation concerns play an even more important role in the public sector than in the private sector. Whether in the public or private sector, reputation concerns have been used to explain, for example, 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.[2]

There are other experiments that investigate how a concern with coming across as well-informed affects behavior. Some experiments deal with individual decision making. Berg et al. (2009) use an experiment to show that an individual decision maker's commitment to a chosen but erroneous course of action is better explained by such reputation concerns than by a concern for consistency per se. Falk and Zimmermann (2017) also find that consistent choices command a higher reputation but can reduce the quality of the decision. Meloso et al. (2023) study a single sender who can only use cheap-talk statements to communicate with an evaluator who assesses the sender after observing the realized state of the world. Their focus is 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. In line with theory, they find that the more uncertain the state is, the more likely the senders are to report truthfully. Contrary to theory, assessments react more to the observed accuracy of statements when senders tend to misrepresent their private information than when senders tend to reveal truthfully.[3]

Like us, Mattozzi and Nakaguma (2023), Fehrler and Hughes (2018), and Fehrler and Janas (2021) study behavior of committees of reputation-concerned decision makers in laboratory experiments, but their focus is different. The first two papers study the effect of secrecy and transparency of the decision-making process on the behavior of subjects and the quality of decisions made. In Mattozzi and Nakaguma (2023), 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. 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 members unwilling to change their minds. Fehrler and Janas (2021) study under which conditions delegation of

a decision to a group of experts yields better decisions for the principal than consulting experts individually. They find evidence for a trade-off between information acquisition by the experts—better under individual consultation by the principal—and information aggregation—better in group decision making.

The rest of the paper is organized as follows. We present the theory of Visser and Swank (2007) on which our experiment is based in the next section and the experimental design in Section 3. We present the results related to game-theoretic predictions in Sections 4 and 5 and the information-theoretic analysis in Section 6. We discuss the findings in Section 7. The Online Appendix contains additional information about the experiment, various robustness checks, and the instructions we used.

2. A Model of Decision Making by Reputation-Concerned Committees

2.1. Setup

The experimental design follows a simplified version of the model of VS. A 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 $\mu \in \{-h, h\}$ and $\Pr(\mu = h) = 1/2$. Thus, μ denotes 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 unconventional.

Stage 0. Nature determines both the state, μ, and the ability level of each committee member $i \in 1, 2$, $a_i \in \{\underline{a}, \overline{a}\}$, with $\Pr(a_i = \overline{a}) = \pi \in (0, 1)$, where $\overline{a}$ stands for high ability and $\underline{a}$ for low ability. These draws are independent. Nature reveals neither the state nor a member's ability to any member or the market. As a result, members and the market only know that each member is of high ability with prior probability π.[4]

Stage 1. Each member receives a private signal $s_i \in \{s^g, s^b\}$ about the state. The quality of s_i depends on a_i. If i is highly able, i receives a high-quality signal that perfectly reveals the state,

$$\Pr(s_i = s^g \mid \mu = h, a_i = \overline{a}) = \Pr(s_i = s^b \mid \mu = -h, a_i = \overline{a}) = 1.$$

If i is of low ability, i receives a low-quality signal that contains no information about the state,

$$\Pr(s_i = s^g \mid \mu = h, a_i = \underline{a}) = \Pr(s_i = s^b \mid \mu = -h, a_i = \underline{a}) = 1/2.$$

As a member is of high ability with probability π, the prior likelihood that a private signal matches the state is $\pi + (1 - \pi)/2 > 1/2$.

Stage 2, deliberation stage. Members simultaneously send cheap talk messages $m_i \in \{m^g, m^b\}$ about their private signals to each other in private.

Stage 3, voting stage. Members simultaneously cast their votes on the project, $v_i \in \{v^1, v^0\}$, where $v_i = v^1$ is a vote for $Y = 1$ and $v_i = v^0$ for $Y = 0$. If both members vote v^1, the committee implements $Y = 1$; otherwise, $Y = 0$ is implemented.

Stage 4, statement stage. Members simultaneously send cheap talk statements ω_i to the market. This statement can be about anything that prevailed in the meeting. Let $\omega = (\omega_1, \omega_2)$.

Stage 5. The market observes Y and ω and determines its assessment of committee member i, $\hat{\pi}_i$, the belief that a member is of high ability. Committee member i's payoff in state μ after decision Y and assessment $\hat{\pi}_i$ equals $Y \cdot (p + \mu) + \lambda \hat{\pi}_i$ with $\lambda \geq 0$ the weight put on the assessment.

2.2. Equilibrium Predictions

Games with cheap talk and voting typically have multiple equilibria, and the game in VS is no exception. We begin by stating the equilibrium that VS study and then discuss other equilibria. Let $\hat{\pi}^1 = [(1 + \pi)/(1 + \pi^2)]\pi$, $\hat{\pi}^0 = [(3 - \pi)/(3 - \pi^2)]\pi$, $\overline{\lambda} = -p/(\hat{\pi}^1 - \hat{\pi}^0)$, and

$$\hat{\pi}^1_{\beta^*} = \frac{(1 + \pi) + 2\beta^*(1 - \pi)}{(1 + \pi^2) + 2\beta^*(1 - \pi)(1 + \pi)}\pi \quad \text{and}$$

$$\hat{\pi}^0_{\beta^*} = \frac{3 - \pi - 2\beta^*(1 - \pi)}{3 - \pi^2 - 2\beta^*(1 - \pi)(1 + \pi)}\pi.$$

Note that $\hat{\pi}^1 > \hat{\pi}^0$ and $\hat{\pi}^1_{\beta^*} > \hat{\pi}^0_{\beta^*}$ for all $\beta^* \in (0, 1/2)$.

Proposition 1 (Visser and Swank 2007). *The following behavior of committee members and assessments form an equilibrium:*

A. *For* $0 \leq \lambda \leq \overline{\lambda}$,

1. *Deliberation stage: each member truthfully reveals the member's private information.*

2. *Voting stage: each member votes* $v_i = v^1$ *if* $(s_i, m_{-i}) = (s^g, m^g)$ *and* $v_i = v^0$ *otherwise.*

3. *Statement stage: for* $\lambda > 0$*, and conditional on* Y*, each member uses a pooling strategy in cheap talk statements; for* $\lambda = 0$*, the statement strategy is undetermined.*

4. *Market assessments: for* $\lambda > 0$*,* $\hat{\pi}_i = \hat{\pi}^1$ *for* $Y = 1$ *and* $\hat{\pi}_i = \hat{\pi}^0$ *for* $Y = 0$ *for all* ω*; for* $\lambda = 0$*,* $\hat{\pi}_i = \hat{\pi}^1$ *for* $Y = 1$ *for all* ω*, and* $\mathbb{E}[\hat{\pi}(0, \omega)] = \hat{\pi}^0$*, where the uncertainty is over* ω*.*

B. *For* $\lambda > \overline{\lambda}$,

1. *Deliberation stage: each member truthfully reveals the member's private information.*

2. *Voting stage: each member votes* $v_i = v^1$ *if* $(s_i, m_{-i}) = (s^g, m^g)$ *and* $v_i = v^0$ *if* $(s_i, m_{-i}) = (s^b, m^b)$*; for* $s_i \neq m_{-i}$*, members vote such that the project is implemented with probability* β^* *satisfying*

$$-p = \lambda(\hat{\pi}^1_{\beta^*} - \hat{\pi}^0_{\beta^*}) \text{ with } \beta^* \in (0, 1/2). \tag{1}$$

where $\hat{\pi}^1_{\beta^*}$ *and* $\hat{\pi}^0_{\beta^*}$ *denote the posterior beliefs following* $Y = 1$

and $Y = 0$, respectively, both in accordance with Bayes' rule and the equilibrium value of β^*.

3. *Statement stage: conditional on Y, each member uses a pooling strategy in cheap talk statements.*

4. *Market assessments: $\hat{\pi}_i = \hat{\pi}^1_{\beta^*}$ for $Y = 1$ and $\hat{\pi}_i = \hat{\pi}^0_{\beta^*}$ for $Y = 0$ for all ω used in equilibrium.*

The proof is in Online Appendix A; the intuition follows here. Because members have common preferences, they share their private signals in the deliberation stage.

Because both members are equally likely to be of high ability, conflicting private signals cancel each other out in the expected project value calculation, $\mathbb{E}[\mu | s^g, s^b] = 0$. Thus, members who want to maximize expected project payoff implement the project only if they receive two positive signals and reject it otherwise.[5] A market that conjectures this relationship between signal pairs and decisions infers from $Y = 1$ that both members received the same, positive signal. The market infers from $Y = 0$ that either both members received the same, negative signal or they received conflicting signals. As two high-ability members receive the same, state-matching signal by construction, if a committee receives conflicting signals, at least one member must be of low ability.[6] With this inference, $Y = 1$ commands a higher reputation than $Y = 0$, $\hat{\pi}^1 > \hat{\pi}^0$.

As a result of the higher assessment following implementation, members face a trade-off in the case of conflicting signals: from a project-value perspective, rejection is best; from a reputation perspective, implementation is best. For a low weight on reputation, $0 \leq \lambda \leq \overline{\lambda}$, a member prefers rejection over implementation as described in part A of the proposition. However, if the weight on reputation is large enough, $\lambda > \overline{\lambda}$, the committee distorts the decision on the projects by implementing it with probability β^* in the case of conflicting signals to gain the stronger reputation. Rational markets see through this inclination, and the equilibrium gain in reputation decreases, $\hat{\pi}^1_{\beta^*} - \hat{\pi}^0_{\beta^*} < \hat{\pi}^1 - \hat{\pi}^0$. In equilibrium, the gain in reputation from implementation in the case of conflicting signals exactly offsets the expected project loss, $-p = \lambda(\hat{\pi}^1_{\beta^*} - \hat{\pi}^0_{\beta^*})$ as described in part B. Finally, the statements members sent to the market are cheap talk and don't affect the decision Y. As a result, if one statement were to lead to a higher reputation than another, a member would always use the former. Thus, in equilibrium, the market ignores these statements. VS draw an additional conclusion that is plausible but not dictated by game-theoretic logic: members show a united front and speak with one voice to the market in support of the decision taken.[7]

When $\lambda = 0$, statements are payoff-irrelevant, and the theory of VS cannot predict which statements will be used in equilibrium. Logically, statements are at least as informative in the case of $\lambda = 0$ as in the case of $\lambda > 0$ as they are completely uninformative in the latter case. We expect that statements are more informative when assessments are not part of committee member payoffs than when they are. This expectation seems warranted on the basis of introspection and other experiments that find that subjects, even those who would benefit from lying or not communicating anything, often prefer to tell (part of) the truth. We return to this phenomenon of overcommunication in the "Discussion" section. The implication is that evaluators are able to glean more information from the statements in the NoA-Stm treatment and should rely on these statements more strongly when determining their assessments than in the A-Stm treatment.

2.2.1. Multiple Equilibria. For each λ, Proposition 1 characterizes two equilibrium relationships, one between signals and the decision on the project—and, thus, expected project payoff—and one between signals on the one hand and decisions and statements on the other—and, thus, reputation payoffs. There are multiple equilibria that lead to the same relationships and, thus, the same payoffs. For example, there are multiple ways that the committee can implement the project with probability β^* when it has received conflicting signals: members can coordinate their votes or use asymmetric voting strategies; the chat in the laboratory facilitates coordination and is often observed. For $\lambda \leq \overline{\lambda}$, the equilibrium relationship can also be implemented without sharing private signals. Thanks to the required unanimous vote, it suffices that a member base the member's vote exclusively on the member's own signal.

In these equilibria, $Y = 1$ commands a higher reputation than $Y = 0$, and this causes a distorted decision if reputation concerns are sufficiently strong. In Online Appendix A.2, we argue that equilibria in which both decisions command the same (expected) reputation are improbable.

In what follows, we, therefore, base our hypotheses on Proposition 1.

3. Experiment: Treatments, Hypotheses, and Procedures

3.1. Treatments and Hypotheses

We use a 2×2 design in which treatments are characterized by the presence or absence of a statement stage (Stm or NoStm) and by the presence or absence of assessments by evaluators in the payoff functions of committee members (A or NoA); see Table 1. In all treatments, we set $\pi = 2/3$, $p = -5$, and $h = 115$. Thus, project implementation yields either 110 or -120. The reputation of a member is implemented as the average assessment that the member obtains from four evaluators. Assessments are measured on a scale from 0% to 100%, indicating the perceived probability that a member is of high ability.

Table 1. Differences Between Treatments

Treatment	Weight λ on assessments	Statement stage?
A-Stm	100	Yes
NoA-Stm	0	Yes
A-NoStm	100	No
NoA-NoStm	0	No

Notes. Description of the differences between treatments. Treatments differ in the weight put on assessments in payoffs of committee members and in the presence or absence of a statement stage.

In the treatments with reputation concerns, we add the member's average assessment to the project payoff obtained by a committee member (i.e., $\lambda = 100$). It can be checked that $\overline{\lambda} = 31.1$ and $\beta^* = 0.33$.

We compare the A-Stm and A-NoStm treatments with the NoA-Stm and NoA-NoStm treatments to establish the effect of reputation concerns on committee behavior and market assessments; we compare the A-Stm and NoA-Stm treatments with the A-NoStm and NoA-NoStm treatments to establish the effect of the cheap-talk channel on committee behavior and market assessments. In what follows, we call the difference in assessment between $Y = 1$ and $Y = 0$ the assessment gap. We say that the assessment gap is positive if $Y = 1$ yields a higher assessment than $Y = 0$.[8] Based on the model, we predict the following:

Evaluators:

EV-1. In all treatments, the assessment gap is positive.

EV-2. The assessment gap is smaller in A-Stm than in NoA-Stm and smaller in A-NoStm than in NoA-NoStm.

EV-3. The assessment gap is as large in A-Stm as in A-NoStm and as large in NoA-Stm as in NoA-NoStm.

EV-4. Assessments respond less positively to statements of confidence in A-Stm than in NoA-Stm.

EV-5. If committee members do not behave according to model predictions, evaluators best reply to observed committee behavior in all treatments.

Committee members:

If committee members act as a single committee and not as two separate individuals, they inform each other about their private signal. Before we test any of the model predictions, we, therefore, check, in all treatments, that members do not lie about the private signals they receive.

CM-1. Voting stage:[9] In all treatments, two positive signals yield $Y = 1$, two negative signals $Y = 0$; in A-Stm and A-NoStm, conflicting signals yield a higher frequency of $Y = 1$ than in NoA-Stm and NoA-NoStm, respectively.

CM-2. Statements: statements contain less information about ability in the A-Stm treatment than in the NoA-Stm treatment.

CM-3. If evaluators do not behave according to model predictions, committee members form beliefs that are consistent with actual assessment formation and best reply to those beliefs.

3.2. Experimental Procedures

We begin by describing the procedures in the A-Stm treatment. At the start of each session and before assigning roles to subjects, we handed out written instructions that covered both roles and went through those instructions verbally. Next, the computer randomly assigned half of the subjects the role of committee member and the other half the role of evaluator.[10] Our schedule of matching subjects needs to balance two goals. The first goal is the avoidance of any uncontrolled dynamic incentives that interfere with the controlled incentives. The second 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 stable, fixed matches.

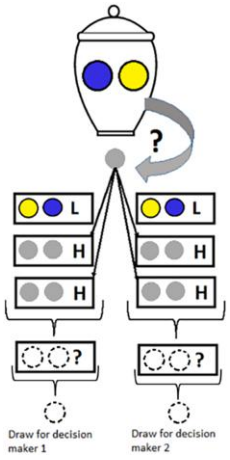
We, therefore, chose an intermediate form. We kept stable throughout the experiment a subject's role (committee member or evaluator), the composition of each committee (consisting of two committee members), and the matching between committees and evaluators. Throughout the experiment the members of two committees were assessed by the same four evaluators. This matching created match groups of subjects.

To prevent the identification of subjects over rounds and, thus, reduce the risk of any uncontrolled dynamic incentives, the software randomly determined the actual evaluators behind the labels evaluator 1–evaluator 4 on each committee member's screen every round. Similarly, on an evaluator's screen, 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 were randomly determined in every round. An additional benefit of fixing the committee composition is the reduction of the time members spend greeting each other and developing a common frame of reference for the experiment.

To explain the relationship between all random variables in the experiment—ability, state of the world, and signals—we designed a novel scheme that summarizes all random draws done in a round. In this figure, quantities of balls and boxes were chosen so that all random draws occurred with equal probabilities. We used it in the written instructions; it also figured prominently on subjects' computer screens.

The scheme starts at the top with a jar containing 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, and the yellow ball the good state. Next, the computer draws a ball, represented by

Figure 1. (Color online) Graphical Depiction of the Relation Between Private Signals on the One Hand and State of Nature and Ability Levels on the Other



an arrow with a question mark below the jar. As the state is not shown to any subject, the color of the ball drawn in the figure is gray. Below the gray ball, there are two columns of boxes, one column for each member of 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 like the jar to indicate that it has no information about the state. Next, the computer randomly selects one of the three boxes, as $\pi = 2/3$, indicated with a downward pointing brace to a box with balls drawn in dashed lines and a question mark rather than a letter. From this box, one ball is drawn at random, again indicated with a downward brace. The color of this ball is a member's private signal. The figure is reproduced in Figure 1.

In each round of 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 receiving their private signals, members could use a chat window for free-form communication within the committee. Communication was private, that is, remained unobserved by any other participant in the experiment. We chose free-form communication for the deliberation to add realism and to obtain data on their thought processes that can be studied to show, for example, 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). The committee's decision was $Y = 0$ unless both members voted for $Y = 1$. On the next screen, members observed both votes cast and the resulting decision. On the same screen, a member was prompted to state the member's degree of confidence in the decision taken by the group. Possible statements were very doubtful, doubtful, neutral, confident, and very confident. We chose these ordered statements over free-form communication because they allow us to directly analyze any effect these statements have on evaluators' assessments.[11] This screen also provided 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 assessments or 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.[12] Once each evaluator had assessed the four members, each member observed the state of nature, the member's committee's decision, the resulting project payoff, and the assessments for that round.

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 Online Appendix B for details. On the results screen, an evaluator observed the evaluator's payoff per committee member and was reminded of the decision of both committees, members' statements, and the evaluator's own assessments. The placement and identity labels of the committees and their members on this screen were the same as on the screen on which the evaluator provided assessments. Across rounds, however, the placement and identities were randomly determined.

Before the actual experiment began, subjects had to answer questions about the payoffs and probabilities to check their understanding of the setup. After all subjects answered all questions correctly, the actual experiment began. 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 stated otherwise. In total, subjects completed 17 rounds. At the end of the experiment, the computer randomly selected four rounds for payment. Earnings for these rounds were added to the show-up fee of €5. After the experiment, subjects filled out a questionnaire about some background characteristics before getting paid in cash and leaving the laboratory. Sessions lasted between one hour and 45 minutes to two hours, including instructions and payment.

The A-NoStm treatment proceeded as the A-Stm treatment with one exception: after members had taken a

decision, they could 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 a situation without strategic interaction between committee members and evaluators. In theory, a zero weight ($\lambda = 0$) on assessments in the objective function of committee members guarantees not only that members' pay is independent of assessments, but also that behavior of committee members is independent of the presence of evaluators. When designing the NoA-Stm treatment, we could not assume that setting $\lambda = 0$ and for the rest proceeding as in the A-Stm treatment would lead to this desired situation. Both the presence of the evaluators in the session and the announcement that evaluators would assess committee members after observing their decisions and statements were likely to influence the behavior of committee members. 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 of committee members, and they were prompted to submit their assessments as in the A-Stm treatment.[13] The incentives for evaluators were the same as in the other two treatments.

Without the interaction between committee members and evaluators, both the duration of the experiment and the expected payment of the committee members in the NoA-Stm would be significantly smaller than in the other treatments if we were to use the same number of rounds. As the expected duration of the session, the show-up fee, and the expected payment are all part of the standard invitation mail at the laboratory that we used, we had to make a choice. We could change the invitation mail sent to the subjects to reflect the shorter duration and different expected payment but keep the number of rounds and pay schedule the same as in the other treatments. However, the characteristics of the invitation to participate in an experiment affect the characteristics of the participants who register for the experiment. The expected payment and show-up fee (Harrison et al. 2009, Gazzale et al. 2013), the type and ambiguity of the information that is provided about the task (Camerer and Lovallo 1999, Gazzale et al. 2013), and the procedure used to register for the experiment (Slonim et al. 2013) are found to affect participants' risk attitude, ambiguity aversion, and social preferences. Alternatively, we could include additional rounds and adjust the number of periods paid out in the experiment to obtain a similar expected payment and duration as in the other two treatments and leave the invitation unchanged. This option has the additional benefit of collecting extra data; the downside is that the larger number of rounds would allow for more learning. We opted for the second alternative and used the same invitation for all treatments. We adjusted the number of rounds to 30 to equalize the duration of the treatments.[14] In Online Appendix J, we show that strategies are qualitatively the same in the first and second half of this treatment.[15]

The NoA-NoStm treatment proceeded as the NoA-Stm treatment with one exception: after members had taken a decision, they could not send statements to evaluators. Thus, evaluators only observed the decision of the committee before they were asked to assess members.

The first three treatments took place in the ESE-econlab at Erasmus University Rotterdam in September–November 2015. We designed the NoA-NoStm treatment and ran it in April and May 2023. All subjects were invited via the econlab subject pool using ORSEE; see Greiner (2004). They were told that the experiment would be about decision making under uncertainty and would last about two hours. The experiments were programmed in php/my-sql and ran on an external server.

In total, 282 subjects participated in the experiment, 88 in the A-Stm treatment, 80 in the NoA-Stm treatment, 56 in the A-NoStm treatment, and 58 in the NoA-NoStm treatment; see Online Table C.1 for an overview of all observations per treatment. Online Table D.1 presents characteristics of the subjects. About half of the subjects are male, and subjects are about 21 years of age. A majority studies economics or business. On average, subjects earned €21.26 in 2015, approximately $28 at the time of the experiments, and €25.32 in 2023 (approximately $27.35 at the time of the experiment) or €20.25 in 2015 euros.[16]

Finally, the chat conversations of the committee members were independently coded by two research assistants according to a common coding scheme.[17] The two sets of coded conversations were compared and differences resolved by the research assistants.

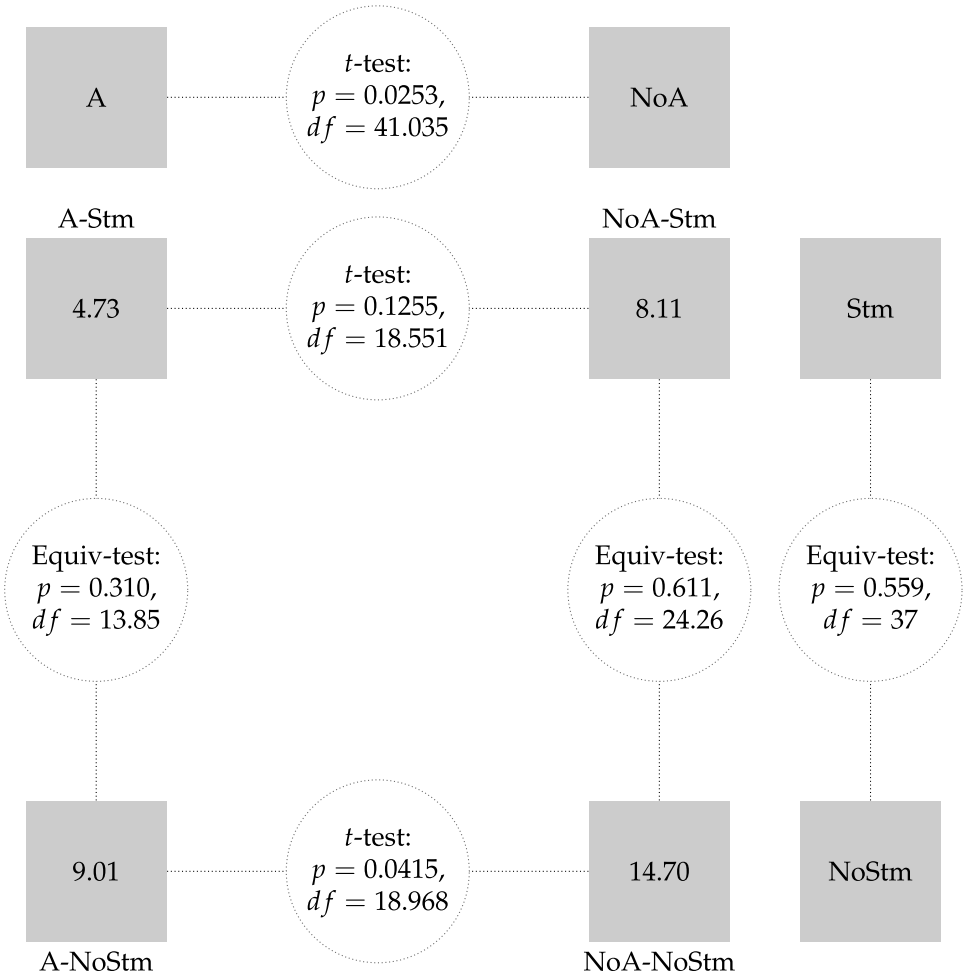
4. Findings: Evaluators

We start by analyzing the determinants and rationality of assessments.

4.1. Assessments (Predictions EV-1 to EV-4)

To investigate prediction EV-1, that the assessment gap is positive, we define an evaluator's assessment gap as the difference in the average assessment that the evaluator gives when the committee decides $Y = 1$ and the evaluator's average assessment when the committee decides $Y = 0$. We aggregate evaluators' assessment

Figure 2. Average Assessment Gaps and *p*-Values for Predicted Differences in Assessment Gaps Across Treatments



Notes. The number in each gray rectangle is the average assessment gap at the match group level in the indicated treatment. Tests are run at match group level and compare match groups across treatments. Degrees of freedom and *p*-values are shown in the circles between the treatments. The horizontal comparisons use a Welch *t*-test to test the null hypothesis of no difference in the assessment gap across treatments with and without assessment entering the payoff functions of committee members against the alternative of the gap being larger in treatments without assessments entering the payoff function of committee members. The vertical comparisons are equivalence tests consisting of a null hypothesis that the difference in assessment gap between treatments with and without statements falls outside the equivalence bounds and the alternative hypothesis that this difference falls inside these bounds.

gaps to the level of independent match groups. Figure 2 reports the average match group assessment gap for each treatment in gray boxes. As theory predicts a positive assessment gap in every treatment, we test the null of no assessment gap against the alternative of a positive gap. The *p*-value of a one-sided *t*-test is 0.0354 for the A-Stm treatment and 0.001 or smaller for the other three treatments.

Result 1 (EV-1). *In every treatment, $Y = 1$ yields a higher assessment than $Y = 0$. Prediction EV-1 is borne out by the data.*

Prediction EV-2 states that the assessment gap is smaller in the treatments with committee members caring about their assessments than without. To investigate this prediction, we test the null of no difference in the assessment gap across treatments with and without committee members caring about their assessments against the alternative of a positive difference across these treatments. We test this with pooled data, A treatments pooled versus NoA treatments pooled, and across individual treatments. The *p*-values of the one-sided *t*-tests at the matching group level are reported in Figure 2.

In the pooled comparison, indicated by the line at the top between A and NoA, $p = 0.0253$. The null of no difference can be rejected at the 1% level: when committee members do not care about assessments, the assessment gap is larger than when they do in line with theory. If we test across individual treatments, the same conclusion can be drawn, but we lose statistical significance in the Stm-treatments because of the smaller sample sizes ($p = 0.1255$ when comparing A-Stm with NoA-Stm and $p = 0.0415$ when comparing A-NoStm with NoA-NoStm).

Result 2 (EV-2). *Treatments with committee members caring about assessments show a smaller assessment gap than treatments without. Prediction EV-2 is borne out by the data.*

EV-3 predicts that the assessment gap is as large in A-Stm as in A-NoStm and as large in NoA-Stm as in NoA-NoStm. The statistically conservative way of investigating a zero-effect prediction (see, e.g., Lakens et al. 2018) is to first formulate a null hypothesis that the effect falls outside some equivalence bounds and an alternative hypothesis that the effect falls inside these bounds. Next, one runs an equivalence test that consists of two one-sided tests to check how the estimated effect compares with the equivalence bounds. If the effect size is statistically unlikely to be larger than the upper bound or smaller than the lower bound, one can reject the null and conclude that the effect is likely not there or relatively small. There is no fixed procedure to select equivalence bounds. Here, we explain ours.

We estimate the (absolute) difference in the assessment gap between the A and NoA treatments to be 3.38 when evaluators receive cheap-talk statements and 5.68 when they do not. By using these effect sizes as bounds, we essentially test whether the effect of having or not having a statement stage is at least as large as the effect of having or not having committee members who care about assessments. We, thus, run two one-sided tests against the largest of the three effect sizes for assessments and select the highest of the two p-values as the relevant p-value for the equivalence test. We do this both with pooled data, Stm-treatments pooled versus NoStm-treatments pooled, and across individual treatments. The p-values of the equivalence tests are reported in Figure 2. In the pooled comparison, indicated by the line between Stm and NoStm on the right, the equivalence tests with a bound 5.68 yields $p = 0.559$, a clear indication that the null of nonequivalence cannot be rejected. If we test across individual treatments, the same conclusion can be drawn albeit at different levels of statistical significance ($p = 0.310$ when comparing A-Stm with A-NoStm and $p = 0.611$ when comparing NoA-Stm with NoA-NoStm).

Result 3 (EV-3). *Treatments with a statement stage do not show an assessment gap that is as large as the assessment gap in treatments without a statement stage. Prediction EV-3 is not borne out by the data.*

Indeed, our results show that the assessment gap is significantly larger in the treatments in which evaluators can base their assessments only on the decision Y.

To investigate prediction EV-4, we use ordinary least squares (OLS) regressions to determine the weights that evaluators attach to the observed decision and statements in their assessments. Table 2 shows that evaluators rarely observe 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, whereas the statements doubtful and very doubtful are grouped together and used as the low-confidence comparison group. Although we formally analyze the statements used when we investigate committee behavior, it is clear from Table 2 that the presence of reputation concerns leads to a boost in the use of the very confident.

In these regressions, we control for round fixed effects because, 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 control for evaluator fixed effects because, first, we want to control for heterogeneity in the response to the scale on which evaluators 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 the way they respond to such scales. Second, these regressions intend to show how differences in observed committee behavior cause differences in assessment, not how average assessments differ across evaluators.

Our schedule of matching committees and evaluators creates match groups. In the A treatments, match groups of evaluators and committee members exclusively deal with each other and share history.

In the NoA treatments, evaluators and committee members do not interact, and we define match groups at each side of the market separately; the pair of evaluators that sees the same two committees share history and form a match group, and the pair of committee members that forms a committee form a match group. We cluster standard errors at the match group level.

Table 3 shows that statements strongly affect how evaluators assess committee members; the coefficients on very confident and confident are considerably larger than the coefficient on $Y = 1$. Given their frequent use, see Table 2, these large coefficients can best be understood to mean that deviating from these common

Table 2. 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.

Table 3. Assessments as a Function of Observables

	Assessment	
Variables	NoStm (1)	Stm (2)
$Y = 1$	14.827***	5.296***
	(2.631)	(1.290)
$Y = 1 \times$ A	−5.085	−2.623
	(2.993)	(2.303)
Very Confident		33.472***
		(3.502)
Very Confident × A		−11.316**
		(4.256)
Confident		22.948***
		(3.355)
Confident × A		−5.024
		(4.262)
Neutral		9.326***
		(2.464)
Neutral × A		−2.428
		(2.673)
Same statement		1.532*
		(0.823)
Same statement × A		0.604
		(1.040)
Observations	5,276	8,040
R^2	0.377	0.525
Adjusted R^2	0.367	0.518
Cluster level	Match	Match
Clusters	31	21
Subject fixed effects	✓	✓
Round fixed effects	✓	✓

Notes. OLS regressions. Assessment is the assessment of ability given by an evaluator on the percentage scale from 0 to 100. $Y = 1$ is a dummy set to one if this members' committee chooses $Y = 1$. A is a dummy variable that is set to one for treatments with assessments entering committee members' payoffs. *Very Confident* is a dummy set to one if this member uses the corresponding cheap-talk statement (similarly for *Confident* and *Neutral*). *Same statement* is a dummy set to one if this member chose the same statement as their partner. Interactions are indicated by ×. Standard errors are in parentheses.

Stars denote significance at the 10% (*), 5% (**), and 1% (***) levels.

statements decreases the assessment. Evaluators seem to believe that members who express doubt in their decision are likely to be of low ability.[18] Group members who use the same statements receive a somewhat higher assessment. Also, the gain in assessment following positive information ($Y = 1$, expressions of confidence) is smaller in the treatment with reputation concerns, A-Stm, than without, NoA-Stm, as can be seen from the signs on the interaction terms.[19]

Result 4 (EV-4). *In A-Stm, assessments respond less to very confident than in NoA-Stm. Prediction EV-4 is borne out by the data.*

4.2. Do Evaluators Make Rational Use of Observed Committee Behavior? (Prediction EV-5)

The fact that prediction EV-3 is not borne out by the data does not necessarily imply that evaluators are biased or irrational. After all, it could be that the presence or absence of a statement stage affects the way that committee members act in the laboratory differently from what theory predicts. For example, we show in Sections 5.2 and 6 that, contrary to equilibrium predictions, cheap-talk statements of many committee members in the A-Stm treatment do contain information about ability. Prediction EV-5 says that the best response of the evaluators takes this into account. In this section, we use 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's (1990, 1998) 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.

We define a variable h_{it} to have a value of 100 if committee member i in round t is of high ability (received high-quality information) and zero if the member is of low ability in round t. This variable captures ability and is defined on the same percentage-point scale as the assessments of the evaluators so that they can be directly compared. We then define the variable *Mistake* as $\Delta_{ijt} = h_{it} - A_{ijt}$. To test for unbiasedness, we calculate the average mistake made by each individual evaluator. We test if this average mistake is different from zero in each treatment using a two-sided t-test in Table 4. The t-tests show that, in the treatments without statements, the average assessment is too low—in A-NoStm by about four percentage points and in NoA-NoStm by about seven percentage points. This corresponds with roughly 7% and 12% of the average assessment, respectively—significant but not extremely large from an economic point of view. In the treatments with cheap-talk statements, differences are considerably smaller and statistically insignificant. Not observing the statements is a handicap to the evaluators.

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. In our experiment, ability is binary, so naively running this test as an OLS against ability might distort the results.

Note that, if information is used efficiently, there should be no systematic link between observed committee behavior and the expected value of the *Mistake* variable. Thus, we study how this variable relates to

Table 4. Assessments and True Ability per Treatment

	Assessment		Ability		Average mistake per evaluator			
Treatment	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	*N*	Pr(mean = 0)
A-Stm	68.23	16.21	67.73	46.76	−0.57	10.26	44	0.7123
NoA-Stm	65.29	17.61	66.90	47.06	1.61	8.38	42	0.2196
A-NoStm	60.04	15.41	64.29	47.93	4.24	9.72	28	0.0288
NoA-NoStm	60.41	18.43	67.74	46.75	7.36	8.94	30	0.0001

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 the unbiasedness of each evaluator.

observable signals in an orthogonality test of the form

$$\Delta_{ijt} = \alpha_0 + \alpha_1 X_{ijt} + \epsilon_{ijt}, \tag{2}$$

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

The information that is available about the two members of a given committee is correlated. Both members have taken the same decision Y and face the same state of nature every round. Similarly, every committee member is evaluated by four evaluators in a matching group, creating a common history within matching groups. Therefore, as in Keane and Runkle (1990, 1998), we cannot assume that the ϵ_{ijt} are independent within rounds or within matching groups. These authors show that clustering standard errors yields a consistent estimate of the variance of the coefficients. 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).[20] This bootstrap procedure estimates error terms within clusters so that we need sufficient variation of the X_{ijt} variables in each cluster. All of our explanatory variables are dummies that have limited variance. Furthermore, both neutral and low confidence statements are relatively rare in the experiment, which forces us to merge these statements in a single bin. We also drop the *Same Statement* variable because of the lack of variation in some clusters.[21]

Column (1) in Table 5 shows no signs of systematic mistakes in A-Stm. In the other treatment with statements, column (2), the coefficient of *Confident* is statistically significant. Its sign is opposite to the constant. To interpret this pattern, note that *Confident* is the modal statement in this treatment (see Section 5.3). If we test whether the sum of the constant and that coefficient is different from zero, we get a nonsignificant result ($p = 0.491$ in an F-test). This coefficient, therefore, seems to indicate that evaluators assess the relatively uncommon statements wrongly. Overall, evaluators tend to make efficient use of available information—decisions and statements—in the treatments with statements.

In the treatments without statements, columns (3) and (4), the constants are significant. This suggests that evaluators, on average, give biased assessments: the average level of the assessments is too low. This is consistent with what we see in Table 4.

Our test has the statistical power to detect systematic deviations from information efficiency. To see this, we run the same regressions with the same clusters but now include a dummy variable, *Confl. Signals*, that is set to one if committee members receive conflicting signals about the state of nature in a particular round. Conflicting signals are a sure sign that at least one member is of low ability. Hence, *Confl. Signals* strongly correlates with ability. As *Mistake* linearly dependents on ability, it is systematically related to *Confl. Signals* as well. Evaluators don't observe the signals received by the committee members, so this information is largely unused by the evaluators by construction. Columns (5)–(8) show that the coefficient of *Confl. Signals* is large and strongly significant in all treatments. This means that these regressions have the power to pick up systematic deviations from information efficiency if there are informative signals in the test.

Result 5 (EV-5). *In treatments with statements, prediction EV-5 is by and large borne out by the data: systematic deviations from the rational use of information only surface for infrequent statements in the NoA-Stm treatment. In treatments without statements, evidence for EV-5 is mixed: assessments reflect the information in the committee decision well but are biased downward on average.*

5. Findings: Committee Members

We now discuss the findings for committee members stage by stage.

5.1. Deliberation Stage

We focus on equilibria in which, when committee members have the same objective function, members share their private information in the deliberation stage. It is, thus, important to check whether this assumption is borne out by communication in the chat.

Table 5. Orthogonality Tests

	Mistake							
	Stm		NoStm		Stm		NoStm	
Variables	A (1)	NoA (2)	A (3)	NoA (4)	A (5)	NoA (6)	A (7)	NoA (8)
$Y = 1$	−3.463	2.329	−1.375	−7.781	−8.200**	−2.166	−5.491	−12.56***
	(4.171)	(3.552)	(3.626)	(5.084)	(3.531)	(4.130)	(3.834)	(4.732)
Very Confident	9.110	2.729			−7.846	−22.92***		
	(6.923)	(5.976)			(9.793)	(7.960)		
Confident	8.714	8.780*			0.370	−14.79***		
	(8.871)	(5.113)			(7.221)	(5.135)		
Confl. Signals					−37.01***	−33.42***	−30.77***	−38.38***
					(12.18)	(11.61)	(10.93)	(12.41)
Constant	−6.882	−4.318	4.893*	10.59**	18.73	26.40***	17.09***	22.92***
	(5.784)	(3.980)	(2.921)	(4.393)	(11.98)	(8.537)	(0)	(0)
Observations	3,000	5,036	1,680	3,596	3,000	5,036	1,680	3,596
R^2	0.005	0.007	0.000	0.006	0.103	0.059	0.084	0.120
Cluster level	Match & Round	Match & Round	Match & Round	Match & Round	Match & Round	Match & Round	Match & Round	Match & Round
Bootstrapped standard error	✓	✓	✓	✓	✓	✓	✓	✓
Fixed effects	—	—	—	—	—	—	—	—

Notes. OLS regressions. *Mistake* is equal to the difference between true ability, h_{ijt}, and the assessment of this ability, A_{ijt}, both on a 100-point scale. $Y = 1$ is a dummy set to one if this members' committee chooses $Y = 1$. *Very Confident* is a dummy set to one if this member uses the corresponding cheap-talk statement (similarly for *Confident*). *Confl. Signals* is a dummy set to one if this members' committee received conflicting signals. Bootstrapped, two-way clustered standard errors in parentheses.

Stars denote significance at the 10% (*), 5% (**), and 1% (***) levels.

Table 6 presents key features of the chat between the reception of the private signal and vote casting.[22] Depending on the treatment, in 25%–44% of all member rounds, a member makes claims that explicitly refer to the private signal the member received, such as "I have a yellow ball."

Two things happen as the experiment progresses. Members tend to shorten their sentences and instead report "yellow" to refer to the color of their ball. And members tend to collapse claims about their private signal and about their desired decision into one.

Table 6. Chat Between Reception Private Signal and Vote Casting: Summary Statistics

	Stm		NoStm	
Percentage of member-rounds with	A	NoA	A	NoA
- message about signal in that round (1)	34.3	38.7	44.5	25.4
- message about color in that round (2)	53.2	64.1	56.2	68.7
- union of (1) and (2)	85.6	98.3	99.0	94.0
- misreporting private information	0.3	0.3	0.7	0.1
- message about vote in that round (3)	49.1	47.6	71.0	52.4
- union of (1), (2), and (3)	87.2	98.9	99.3	94.0

Notes. Summary of topics of discussion in chat about private signal and casting of vote, incentivized rounds only. All the messages a committee member sends in the chatbox in a round are treated as a single observation. To determine whether a signal was misreported, we compared the actual signal a subject receives with the variables "message about signal in that round" and "message about color in that round."

Thus, in 53%–68% of all member rounds, they make a statement such as "yellow" without indicating whether this refers to their private signal, to the vote they intend to cast, or both. The union of member rounds with either type of statement covers nearly all member rounds in A-NoStm, NoA-Stm, and NoA-NoStm; in A-Stm, the percentage is high, at 85.6%. In these unions of member rounds, the stated color is rarely in conflict with the actual color of the ball they received: private information is misreported, whether erroneously or to deceive, in less than 1% of all member rounds.

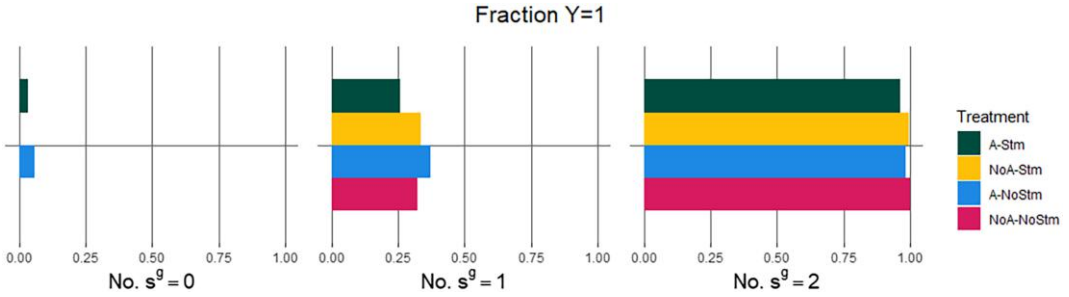
The high percentages of shared signals in combination with the near absence of misreporting means that, by and large, the focus on truthful committee communication is justified.

5.2. Voting Stage (Prediction CM-1)

In the first few rounds, once members have exchanged their private signals, they discuss what decision to take. As expected, most committees quickly agree that, if they receive the same signal, they should vote in line with those signals. In case of conflicting signals, discussions about what to vote remain common. Table 6 shows that the percentage of member rounds with messages about votes lies between 50% and 73% in the various treatments.

Figure 3 shows the fraction of $Y = 1$ decisions within each treatment and for a given number of positive

Figure 3. (Color online) Relationship Between Number of Positive Signals and Committee Decision



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

signals s^g (or yellow balls) that a committee received. As predicted by theory for all treatments, committees choose $Y = 0$ after two negative signals and $Y = 1$ after two positive signals.

But the predicted frequency of distorted decisions—$Y = 1$ in case of conflicting signals—is not found in the data. Not only do committees in NoA treatments distort the decision, they do so with a frequency that is comparable to committees in A treatments.[23]

As in Section 4.1, statements are not expected to have an effect. Therefore, the conservative test of the effect of statements on the amount of distortion in the decisions is an equivalence test. In that section, we used the size of the reputation effects to determine the size of the upper and lower equivalence bound. We follow the same approach here. We set the upper and lower bounds of a proportion test between the A treatments and NoA treatments such that the test narrowly accepts the H_0 of nonequivalence at the 5% significance level. At that equivalence bound, the equivalence test for the Stm and NoStm treatments rejects the H_0 of nonequivalence ($p = 0.0240$). In the distortion of the decisions, the effect of statements appears smaller than that of reputation concerns.[24]

Result 6 (CM-1). *In line with prediction CM-1, committees choose $Y = 0$ after two negative signals and $Y = 1$ after two positive signals in all treatments. In case of conflicting signals, committees distort the decision with a frequency that is comparable across treatments. This null result contradicts CM-1's prediction that reputation concerns lead to more distortions in case of conflicting signals.*

As observed behavior when private signals are conflicting differs from the equilibrium prediction, we analyze in Section 5.4 whether this behavior is optimal given the beliefs that members have about what determines their assessments.

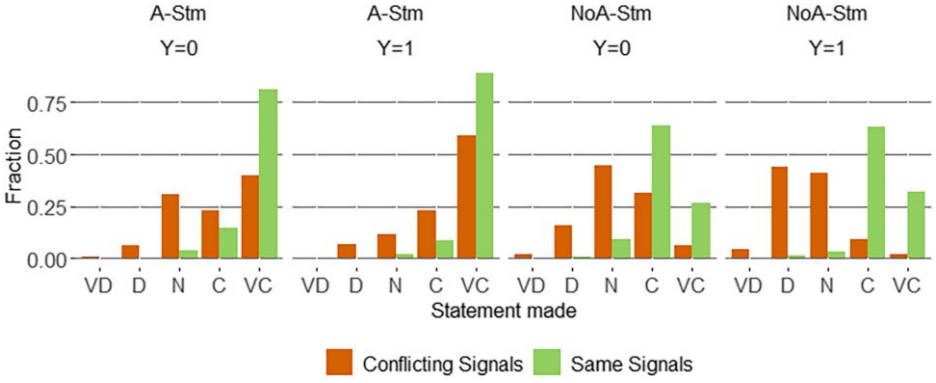
5.3. Statement Stage (Prediction CM-2)

This section only applies to the two treatments with a statement stage. CM-2 predicts that statements are less informative about ability in the A-Stm treatment than in the NoA-Stm treatment.

Figure 4 shows the fraction of member rounds in which a statement is used, conditional on the pair of private signals, the committee decision, and the treatment. To ensure that the subjects are aware of the signal of their partner when they send a statement, we drop the periods in which the partner's signal is not discussed. A comparison of the left two panels (A-Stm) with the right two (NoA-Stm) shows that concern with assessments makes very confident by far the most common statement. Independent of the actual pair of private signals, expressions of doubt practically disappear when assessments play a role. This difference in distributions across the two treatments is significant for both conflicting and same signal rounds ($\chi^2, p < 0.001$ in both cases). Further comparisons show that the main difference between statements is caused by the presence or absence of conflicting signals and less so by the decision taken. Although a comparison of the statement distributions in same signal rounds (green bars) between rounds with $Y = 1$ and $Y = 0$ shows a difference in the NoA-Stm (last two panels, $\chi^2, p = 0.0097$), there appears to be no difference in the A-Stm treatment (first two panels, χ^2, $p = 0.1267$). However, in both treatments and for both decisions, rounds with conflicting signals (red bars) have different statement distributions than rounds with concurring signals (green bars) ($\chi^2, p < 0.001$ in all four comparisons).

To formally investigate whether the statements indeed contain less information in A-Stm than in NoA-Stm, we calculate the amount of information a single statement has about the presence or absence of conflicting signals. First, note that our χ^2 tests indicate that the main difference in statement distributions is between conflicting and concurring signals. Therefore, we pool the $Y = 1$ and $Y = 0$ rounds and test how difficult it is to distinguish the distribution of statements used in conflicting signal rounds from the distribution used in concurring signal rounds. We view the statement that a member uses when members receive the same signal as a random

Figure 4. (Color online) 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.

variable V with possible statements $x_1, \ldots, x_4$.[25] The associated probabilities $p_1, \ldots, p_4$ are set at the member's empirically observed relative frequencies of the statements. Similarly, we view the statement that a member uses when members received different signals as a random variable W with associated probabilities $q_1, \ldots, q_4$. We then calculate the Jensen–Shannon divergence (JSD) of these distributions of statements.[26] This measure can theoretically go from zero, meaning the distributions are indistinguishable, to one, which implies that a single draw identifies the distribution perfectly:

$$\mathrm{JSD}(V,W) = \frac{1}{2}\sum_i \left(p_i \log_2\left(\frac{p_i}{p_i + q_i}\right) + q_i \log_2\left(\frac{q_i}{p_i + q_i}\right)\right). \tag{3}$$

Figure 5 presents the distribution of members' Jensen–Shannon divergence, using bins of 0.05 width. The modal bin in either treatment, attained by 55% of committee members in the A-Stm treatment and by 39% in the NoA-Stm treatment, is the lowest, [0, 0.05]. As a result, 55% of committee members in A-Stm make it (nearly) impossible for evaluators to glean information about the signal pair they have received from the statements they use. Because a higher value of JSD means that the conflicting signal distribution and the concurring signal distribution are easier to distinguish, we compare the distribution of the JSD between the treatments. We find that the median of the JSDs is higher in NoA-Stm treatment (one-sided exact Wilcoxon rank-sum test, $p = 0.009$). The committee members, thus, reveal more information about their signals in the NoA-Stm treatment than in the A-Stm treatment as predicted in CM-2.[27]

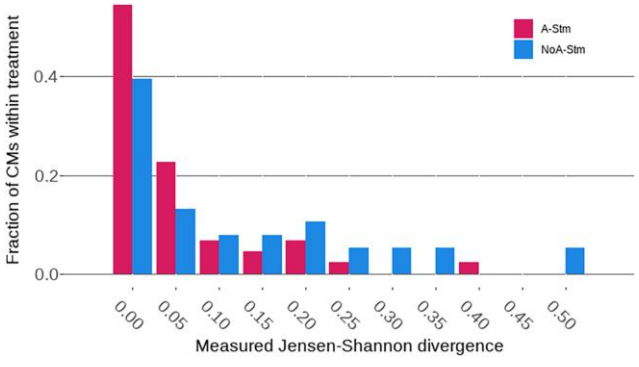
Result 7 (CM-2). *The median committee member in A-Stm reveals less information than the median member in NoA-Stm about ability through the cheap-talk statements. Prediction CM-2 is borne out by the data.*

5.4. Do Committee Members Best Reply to Assessments? (Prediction CM-3)

As equilibrium predictions are only partially borne out by the data, we now investigate the more basic prediction that members' observed behavior is a best reply to the incentives provided by the experimental treatment (prediction CM-3).

For the two NoA treatments, in which committee members' payoffs equal project payoff, the theory predicts the absence of distorted decisions as decision makers are assumed risk neutral. The observed distorted decisions in either treatment could be due to a risk-loving attitude. Committees in any treatment commonly discuss whether to take risk. Online Table D.3 shows that higher degrees of self-reported risk tolerance indeed raise the probability of voting for $Y = 1$ in the case of conflicting signals in all treatments. This

Figure 5. (Color online) Distribution of Jensen–Shannon Divergence for Statements



Notes. The figure shows for both treatments with a statement stage the empirically observed distribution of members' Jensen–Shannon divergence, using 0.05 bins. A member's Jensen–Shannon divergence measures the difference between the empirically observed distribution of statements when committee members received the same signals and when they received conflicting signals.

suggests that the distorted decisions in the NoA treatments are not necessarily the result of a behavioral bias.

Whether committee members in the A treatments best reply depends on how evaluators' assessments respond to the observed behavior of committee members. We ran a regression similar to the one reported in Table 3, column (2), using only the data of treatment A-Stm. The estimated coefficient of $Y = 1$ equals 2.43, and we cannot reject the hypothesis that it is equal to five (F-test, $p = 0.2094$). The gain in reputation just about compensates the expected loss from distorting the decision in this treatment.[28] Committees are then approximately indifferent between $Y = 1$ and $Y = 0$ and can rationally choose either decision. Committee members who fail to use very confident do leave money on the table.

In the A-NoStm treatment, the assessment gap more than compensates the expected loss from distorting the decision.[29] The best reply is then to always choose $Y = 1$ in case of conflicting signals rather than only now and then.

Result 8 (CM-3). *Prediction CM-3 is not borne out by the data. On average, committee members in the A treatments do not best reply to the incentives provided by evaluator assessments: in the A-Stm treatment, they should have made more use of the very confident statement; in the A-NoStm, they should have chosen $Y = 1$ with probability one in case of conflicting signals. In the NoA treatments, the difference between predictions and observed behavior can partly be explained by risk tolerance.*

6. Entropy

The orthogonality tests presented in Section 4.2 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 observe about ability in a treatment. As a result, they cannot be used to study how much of the available information about members' ability is transformed by members' behavior and how much of that information eventually finds its way into the assessments. However, such comparisons are highly relevant; they shed light on how the presence or absence of reputation concerns and cheap talk determine the amount of information on which the reputation market can base its assessments. Without information on ability, the reputation market cannot function as an institution for selection and control of experts. 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 on the cardinal entropy scale. We measure the degree to which the initial uncertainty about a member's ability is reduced by the presence or absence of conflicting signals within the committee and by the observed behavior of committees. We also measure how much of that reduction of information finds its way into the assessments that evaluators provide. Because of the cardinal scale, we can make sensible comparisons across treatments and establish, for example, 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. The entropy of variable X is defined as the expected information in observations of X,

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

A binary variable—for example, 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. 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.[30]

We want to establish how much the initial entropy concerning ability is reduced by 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 measure the reduction in entropy of X (ability) thanks to the observation of Z (conflicting signals, decisions, assessments):

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &= \sum_{i,j} p(x_i, z_j) \log_2 \left(\frac{p(x_i, z_j)}{p(x_i)p(z_j)} \right). \end{aligned} \tag{5}$$

I is called the mutual information of X given Z. One can use this measure to establish how much easier it becomes for a member to determine the member's ability level once the member has observed the signal pair that the member's committee has obtained or for an evaluator to predict a member's ability once the evaluator has observed a decision and, depending on the treatment, the statements.

Table 7 shows 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).[31] The binary variable *Confl. Sign.* is equal to one if committee members receive conflicting signals and equal to zero if they receive the same signals. This is the only information about the ability the computer reveals during a round. The binary variable $Y = 1$ refers to the decision that a group takes, whereas the *Stm4* variable captures the statements used by a committee member

Table 7. Entropy and Mutual Information of Ability Given Various Variables

Treatment	Entropy (1) *Ability*	Mutual information of ability given various variables (2) *Confl. Sign.*	(3) *Info Set*	(4) $Y = 1$	(5) *Stm4*	(6) *Assessment*
A-Stm	0.9075	0.0941	0.0536	0.0002	0.0240	0.0050
NoA-Stm	0.9161	0.0936	0.0910	0.0109	0.0499	0.0228
A-NoStm	0.9407	0.0706	0.0050	0.0050	—	0.0011
NoA-NoStm	0.9074	0.0890	0.0038	0.0038	—	0.0001

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 of ability. The other columns list the estimated mutual information of ability and the respective variables listed at the top of the columns. *Confl. Sign.* is a dummy set to one if the committee received conflicting signals about the state of nature. Y =1 is a dummy set to one if the 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 $Y = 1$ and the *Stm4* variables of both committee members in a single categorical variable with $2 \times 4 \times 4$ categories. *Assessment* captures the assessment given by evaluators in a treatment, transformed to a discrete variable by binning the assessments in five percentage point bins. Online Table L.1 reports the bootstrapped standard errors of these estimates.

and observed by the evaluator. As before, we bin the lower two statements to have enough observations. *Info Set* is a variable that combines $Y = 1$ and the *Stm4* variables of both committee members in a committee,with 32 possible values. *Assessment* captures the assessments given by the evaluators in treatment-specific bins with a bin containing approximately 5% of the assessments in the treatment.[32]

In all treatments, the estimated entropy of members' ability is of similar size; see column (1). The pair of private signals that members receive significantly reduces the entropy; see column (2), but considerable uncertainty remains.

As to the information embodied in committee member behavior that evaluators observe, column (3) shows that, in the two treatments with the statement stage, committee members reveal considerably more about their ability than in the two treatments without that stage irrespective of whether members care about their assessments or not. This difference in mutual information because of the presence or absence of a statement stage is much larger than the difference because of the presence or absence of assessments in members' payoffs.

A comparison of columns (4) and (5) shows that cheap-talk statements contain considerably more information than the decision, a costly signal: nearly five times more in the absence of a concern with assessments and much more in the presence of such concerns.

Only a limited amount of information about ability is available to the committee members themselves (compare columns (1) and (2)), and then, the available information is largely concealed by the behavior of subjects (compare columns (2) and (3)). Consequently, when we turn to the mutual information of ability and assessments, column (6), we see that the amounts are low in all treatments.

To sum up, in the absence of statements, the task of evaluators to assess committee members' ability is especially arduous whether committee members care or not about their assessments. In this experiment, words speak louder than costly actions. This justifies the dependence of assessments on these statements that we find in Section 4.1. But, even with statements as a possible source of information, the presence of a concern with assessments considerably reduces how much evaluators can infer from observed committee behavior.

7. Discussion

The aim of the experiment was to answer two sets of questions on the behavior of committee members and the assessments of evaluators using a laboratory experiment with a 2×2 treatment design. We find mixed evidence for the various predictions that we derived from the theory of VS. We first summarize this evidence and then discuss some of its implications.

Before summarizing the evidence on predictions that require comparisons across treatments, we begin with a prediction that applies to every treatment, the existence of a positive assessment gap: evaluators are predicted to give higher assessments after committees decide to implement the project ($Y = 1$) than after they decide to reject it ($Y = 0$). This prediction, EV-1, is borne out by the data. This is an important finding: even though the quality of the decision cannot be established as the state of nature remains unobserved, human evaluators do update their beliefs about a decision maker's ability depending on the decision that the committee made; indeed, in line with theory, they think that a committee member is more likely to be able to take the correct decision after implementation than after rejection.

As to the first set of questions on the effect of reputation concerns on committee members' behavior and the quality of evaluator assessments, we find that committees distort the decision as much in the A treatments as in the NoA treatments. This null result is a deviation from theory (prediction CM-1). On the other hand, we do find that committee members make it harder to infer

information about ability from their statements when they care about their reputations than when they do not in line with prediction CM-2. Turning to evaluators, we find that the assessment gap is smaller in treatments with committee members caring about assessments than in treatments without in line with EV-2. We also find that evaluators pay less attention to the statement very confident when committee members care about their assessments in line with EV-4.

As to the second set of questions on the effect of the presence or absence of cheap-talk statements on committee decisions and evaluator assessments, we find a null effect on committee decisions; this null effect is in line with theory. On the other hand, the assessment gap is larger in treatments without statements than with, contrary to equilibrium prediction EV-3.

The mixed support for the equilibrium predictions leads to the question whether subjects best reply to each others' observed behavior. When we analyze whether evaluators rationally base their assessments on observed behavior of committees using an orthogonality test, we find that, in treatments with cheap-talk statements, assessments are quite rational; in treatments without, they are biased (on average, too low). This holds whether assessments enter committee members' payoffs or not (EV-5). When we analyze whether committees best reply to expected project payoffs and, in treatments with assessments, to observed assessments, we find that committees fail to use some of the possibilities to raise expected payoffs in all treatments (CM-3). In treatments without assessments, committees could have earned a higher expected payoff if they had implemented the project in the case of conflicting signals less often. In the treatments with assessments, committee members could have earned higher assessments by using the most optimistic cheap-talk statement very confident more often (in the A-Stm treatment) or by implementing more often the project in case of conflicting signals (in the A-NoStm treatment).

Given that, on average, evaluators best reply in Stm treatments but not in NoStm treatments, whereas committee members significantly deviate from their best replies in all treatments, we continue the discussion of the findings for the two types of subjects separately.

Evaluators face a difficult problem as they do not observe the state of the world when assessing committee members. Their assessments can only be based on committee decisions and, in the Stm treatments, on cheap-talk statements. We find that the quality of the assessments—their rationality given the available information as revealed by the orthogonality test and the average mistakes and the quantity of information about the true ability that assessments contain as measured by mutual information—is higher when evaluators have access to cheap-talk statements. That the additional source of information does not confuse evaluators but helps them in improving the quality of assessments is important to note. Equally important is the finding that with cheap-talk statements, the assessments are rational also in the A-Stm treatment. This holds even though committee members use the statements strategically to obtain higher assessments.

When it comes to evaluators, the bottom line is that the quality of their assessments is hurt by the absence of cheap-talk statements rather than by the presence of committee members who strategically seek to influence them. The functioning of the reputation market as an institution that selects and motivates able experts (Fama 1980, Holmström 1999), however, not only depends on the quality of assessments, as quality is defined relative to the amount of information that is available to evaluators, but also on the absolute amount of information used in the assessments. Both the absence of cheap-talk statements and the presence of reputation concerns reduce the amount of information to which evaluators have access. Here too, we find that the absence of statements is more damaging than the presence of reputation concerns.

Committee members can raise expected payoffs in all treatments. We argue in Section 5.4 that, in the NoA treatments, members' risk tolerance may partly explain why, contrary to the equilibrium prediction, they make the risky choice with a negative expected payoff when they have received conflicting signals. We find that committees distort the decision to a comparable degree in the A and NoA treatments. In treatments with statements, this can be partly explained by information about ability being communicated via the cheap talk statements. When committee members have access to both a costly decision and a cheap talk statement to communicate about their ability, the cheap talk is, by definition, cheaper. Therefore, committee members likely prefer to manage their reputation using cheap-talk statements rather than costly decision distortions. Consistent with this idea, we show in Section 6 that cheap talk communication contains 5 to 10 times more information about ability than decisions. This reduces the pressure to distort decisions in the A-Stm treatment, and so, on the one hand, committee behavior in that treatment moves closer to behavior in the NoA-Stm treatment. On the other hand, we notice that conflicting signals lead to discussions about the decision payoffs: maximizing the expected payoff means choosing $Y = 0$, but by doing so, members exclude the chance of receiving a positive decision payoff. Online Table M.4 shows that this zero-payoff dilemma is especially felt in the NoA treatments because decision payoffs are the only payoffs they receive. This increases the pressure to distort in NoA treatments and moves committee behavior in those treatments closer to behavior in the A treatments.

It could be that committee members need to gain experience to learn how to play the game. In Online Appendix J, we compare the behavior of committee

members across the first and second half of the experiment. As far as statements are concerned, we find that confident statements do become more common in the second half than in the first half of the A-STM treatment, but not in the NOA-STM treatment.[33] In none of the treatments do we find evidence that voting behavior differs across the two halves. This does not suggest that subjects learn to play the equilibrium within the time span of the experiment. In particular, we find no evidence that, in the A-STM treatment, committee members replace cheap-talk statements by the decision on the project as their channel to signal ability.[34]

When it comes to committee members, the bottom line is that further research, both experimental and theoretical, on the combination of costless and costly messages for information transmission seems worthwhile. We noticed, in the introduction, that the combination of costly signals and cheap talk is common in practice. We find that, on average, there is less information in the costly decision than in the cheap-talk statements. In another experiment, De Haan et al. (2015) also find that, even when costly signals are available, experimental subjects prefer to communicate through cheap talk. In the theoretical literature on such combinations, the focus has been on the role that is left for cheap talk if a costly signal becomes available (Austen-Smith and Banks 2000, Karamychev and Visser 2017). The most common result is that the cheap-talk signal contains little to no information. Experiments on cheap talk show that it conveys more information than theoretically expected. This phenomenon has been called overcommunication. The focus has been on contexts in which senders can tell the truth or lie about a privately received signal.[35] In our experiment, committee members can claim to be very confident in the decision even though they received conflicting signals. This is different from directly lying about one's private signal. This difference makes the statements more "malleable to ex-post interpretation as truths," to cite Turmunkh et al. (2019, p. 4795), and this appears to reduce lying costs in our experiment as well.

When committee members care about their reputations and can send statements to evaluators, VS stress a specific pooling statement strategy, a united front. It results from the conscious choice to act in tandem, not a coincidentally appearing equality of statements. This is clear from VS's use of the phrase by Frederick H. Schultz, a former governor and vice-chairman of the Federal Open Market Committee: "We should argue in the Board meetings but close ranks in public" (Visser and Swank 2007, p. 339) to illustrate a united front. A proper test of this part 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 brings up the importance of a united front in a Schultz-like manner in the chat.[36] By this test, we find little support for a united front: only two committees pass it.[37]

Acknowledgments

The authors thank two anonymous referees and an associate editor for their detailed comments and their patience over the Covid period and Sebastian Fehrler, Chaim Fershtman, Sacha Kapoor, Debrah Meloso, Lúis Santos-Pinto, Karl Schlag, and Otto Swank as well as seminar audiences at Erasmus University Rotterdam and Middlesex University as well as at the universities of Mannheim, Konstanz, Lausanne, Milan, Vienna, and Delft for comments and discussions. Annikka Lemmens, Erik van Goudoever, Xiaomeng Chen, and Mohamed Orban provided diligent research assistance. A previous version of this paper was called "Committees of Experts in the Laboratory."

Endnotes

[1] Hermalin and Weisbach (2017) review some empirical work that uses observational data that exploits either intertemporal patterns of a manager's compensation that can be explained by career concerns or focuses on industries in which market-based incentives can be measured. Meade and Stasavage (2008), Swank et al. (2008), and Hansen et al. (2017) exploit a change in the publication requirements at the Federal Open Market Committee, the monetary policy committee of the U.S. Federal Reserve System, to empirically investigate the role played by reputation concerns.

[2] 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 (2023). In these papers, a decision maker is concerned with perceived ability (and possibly a state-dependent project payoff). Other concerns may exist; see, for example, Gradwohl (2018) for voters who, besides caring about a state-dependent payoff, care about strategic ambiguity.

[3] Other experiments, such as 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.

[4] VS show that their main findings continue to hold when members know their ability.

[5] For the model to be of interest, parameter values are such that, from a project-value perspective, the project should be implemented if $(s_1, s_2) = (s^g, s^g)$.

[6] Note that assessments depend on the conjectured pairs of signals that lead to a decision, and prior beliefs π are independent of i. Because votes and signals of individual members are not revealed, assessments are the same across members.

[7] Game theory does dictate then that the market should be able to assess a member in the out-of-equilibrium event that the committee was not to show a united front. It is consistent with the model to assume that disagreement leads to a drop in assessment.

[8] In treatments A-STM and NOA-STM, in which evaluators receive cheap-talk statements, the assessment gaps are based on assessments for Y averaged over the cheap-talk statements received.

[9] As there are many voting strategies that yield the same equilibrium relationship between signals and the decision, the focus is on the predicted relationship between signals and the decision.

[10] In the experiment, committees were called groups, committee members decision makers. The participants that form the reputation market are called evaluators.

[11] Statements about confidence in the decision seem more natural than statements about concurring or conflicting signals. Also, in reality, experts may be ambiguous. Markets or committee watchers have to infer the essence from experts' statements. The experiment allows us to include a little bit of this richness. Moreover, we wanted to avoid that we would immediately indicate what subjects—both committee members and evaluators—should look for by limiting statements to a dummy variable about conflicting signals. We can only speculate about the possible consequences of making evaluators base their assessments on free-form communication by committee members rather than structured statements. However, it would certainly lead to substantial measurement error as we would have to translate the communication into variables that can be used in econometric analysis.

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

[13] A similar procedure is used in Falk and Zimmermann (2017) to generate treatments with and without reputation.

[14] To obtain a similar expected payment, we adjusted the number of periods paid out to 10.

[15] As a further check on the effects of duration, we 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 Online Appendix K. In the analysis, we, therefore, merge this data with the rest of the A-Stm treatment.

[16] To convert to 2015 euros, we use the consumer price index provided by Statistics Netherlands, at https://www.cbs.nl/nl-nl/cijfers/detail/83131NED.

[17] See Online Appendices M.2 and M.3 for the coding scheme.

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

[19] A joint test on the significance of all the interaction terms yields $p = 0.0027$.

[20] For details about this clustering, see also Cameron et al. (2011). This combination of two-way clustering and bootstrapped standard errors was run using Stata 18 and the CGMwildboot package obtained from the website of Jonah Gelbach.

[21] Full regressions with all variables with nonbootstrapped standard errors are shown in Online Appendix H. They yield the same results qualitatively but with smaller estimated standard errors.

[22] Online Table M.4 presents a complete overview of all coded variables. In that appendix, we also explain to which variables the various lines in Table 6 correspond.

[23] In a two-sided test of proportions of the pooled A treatments and pooled NoA treatments, $p = 0.606$.

[24] The equivalence tests use two one-sided proportion tests to see if the proportion of rounds with $Y = 1$ in rounds with conflicting signals in the Stm (A) treatment are equivalent to the same proportion in the NoStm (NoA) treatment. The lower and upper bounds are set so that the ratio between the lower (upper) bound and the reverence proportion are equal to 1/1.34 (1.34). This yields $p = 0.0501$ for the comparison between A and NoA and $p = 0.0240$ for the comparison between Stm and NoStm.

[25] We use four possible statements and pool the observations with very doubtful and doubtful because of their low usage.

[26] The Jensen–Shannon divergence is based on measures of entropy from information theory. The seminal paper on entropy in information theory is Shannon (1948). A textbook presentation can be found in Luenberger (2006). A recent discussion of the link between entropy and the value of information in a decision problem can be found in Frankel and Kamenica (2019).

[27] We also ran the test separating the rounds with $Y = 1$ and $Y = 0$. As a result, we differentiate rounds based on the treatment, statements made, conflicting signals, and decisions. The resulting four-dimensional split has low numbers of observations in some cells (in particular, low confidence statements in $Y = 1$ rounds in A-Stm), which reduces the power of the tests. To run these tests, we, therefore, have to combine neutral with the doubtful and very doubtful statements. Results are similar, but only the difference in the case of $Y = 0$ remains statistically significant (one-sided exact Wilcoxon rank-sum test, $p = 0.04$).

[28] A similar test using the assessment gaps in Figure 2 yields the same conclusion (two-sided t-test, $p = 0.8133$). However, this assessment gap does not take into account the effect of the statements.

[29] In a regression using only data from the A-NoStm treatment, we find a coefficient on $Y = 1$ equal to 9.6. An F-test rejects the hypothesis that this coefficient equals five ($p = 0.0336$).

[30] We use the Jensen–Shannon divergence in Section 5.3. It can be expressed in terms of the entropy of the random variables, $\mathrm{JSD}(V, W) = H((V + W)/2) - (H(V) + H(W))/2$.

[31] As regression techniques to estimate entropy and related variables are unavailable, we use an estimate based on maximum likelihood to determine the empirical entropy. We use a bias correction term known as a Miller–Madow bias correction because the maximum likelihood estimate of entropy is biased even asymptotically. See Paninski (2003) for details. Calculations were all done with the infotheo package in *R* of Meyer (2014).

[32] Assessments could fall anywhere in the range 0–100, whereas ability is binary. Information-theoretic measures do not take the meaning of the variables into account: an observation with an assessment, say 51%, that is only given to a single high-ability committee member is treated as a perfect statistical flag for high ability. This is clearly not what was meant by the evaluators; it is also independent of the ability of the average committee member that received a 50% or 52% evaluation. Such perfect statistical flags strongly influence the estimation of mutual information. We prevent this by assigning assessments in each treatment to 20 bins with equal numbers of observations.

[33] The latter finding should allay concerns that the additional rounds aided learning.

[34] Online Appendix J also investigates evaluators' assessments. We find no evidence that evaluators react differently to the observed behavior of the committees over the two halves.

[35] See Dickhaut et al. (1995) and Cai and Wang (2006) for single-agent decision problems, Goeree and Yariv (2011) and Fehrler and Hughes (2018) for a committee setting, and Blume et al. (2020) for a survey. Meloso et al. (2023) find both overcommunication and undercommunication.

[36] We prefer this test over one based on a sentence—common in the A-Stm treatment—such as "Shall we choose confident?" as it lacks an articulation of the importance of using the same statement.

[37] See excerpts 4 and 5 in Online Appendix M.1.

References

Austen-Smith D, Banks JS (2000) Cheap talk and burned money. *J. Econom. Theory* 91(1):1–16.

Berg JE, Dickhaut JW, Kanodia C (2009) The role of information asymmetry in escalation phenomena: Empirical evidence. *J. Econom. Behav. Organ.* 69(2):135–147.

Blume A, Lai EK, Lim W (2020) Strategic information transmission: A survey of experiments and theoretical foundations. Capra CM, Croson RTA, Rigdon ML, Rosenblat TS, eds. *Handbook of Experimental Game Theory* (Edward Elgar Publishing, Cheltenham), 311–347.

Butler JV, Miller JB (2018) Social risk and the dimensionality of intentions. *Management Sci.* 64(6):2787–2796.

Cai H, Wang JT-Y (2006) Overcommunication in strategic information transmission games. *Games Econom. Behav.* 56(1):7–36.

Camerer C, Lovallo D (1999) Overconfidence and excess entry: An experimental approach. *Amer. Econom. Rev.* 89(1):306–318.

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. *Rev. Econom. Statist.* 90(3):414–427.

Cameron AC, Gelbach JB, Miller DL (2011) Robust inference with multiway clustering. *J. Bus. Econom. Statist.* 29(2):238–249.

Chierchia G, Nagel R, Coricelli G (2018) Betting on nature or betting on others: Anti-coordination induces uniquely high levels of entropy. *Sci. Rep.* 8(1):1–11.

De Haan T, Offerman T, Sloof R (2015) Money talks? An experimental investigation of cheap talk and burned money. *Internat. Econom. Rev.* 56(4):1385–1426.

Dewatripont M, Jewitt I, Tirole J (1999a) The economics of career concerns, part 1: Comparing information structures. *Rev. Econom. Stud.* 66(1):183–198.

Dewatripont M, Jewitt I, Tirole J (1999b) The economics of career concerns, part 2: Application to missions and accountability of government agencies. *Rev. Econom. Stud.* 66(1):199–217.

Dickhaut JW, McCabe KA, Mukherji A (1995) An experimental study of strategic information transmission. *Econom. Theory* 6(3):389–403.

Falk A, Zimmermann F (2017) Consistency as a signal of skills. *Management Sci.* 63(7):2197–2210.

Fama EF (1980) Agency problems and the theory of the firm. *J. Political Econom.* 88(2):288–307.

Fehrler S, Hughes N (2018) How transparency kills information aggregation: Theory and experiment. *Amer. Econom. J. Microeconomics* 10(1):181–209.

Fehrler S, Janas M (2021) Delegation to a group. *Management Sci.* 67(6):3714–3743.

Frankel A, Kamenica E (2019) Quantifying information and uncertainty. *Amer. Econom. Rev.* 109(10):3650–3680.

Furnham A, Boo HC (2011) A literature review of the anchoring effect. *J. Socio-Econom.* 40(1):35–42.

Gazzale R, Jamison J, Karlan A, Karlan D (2013) Ambiguous solicitation: Ambiguous prescription. *Econom. Inquiry* 51(1):1002–1011.

Goeree JK, Yariv L (2011) An experimental study of collective deliberation. *Econometrica* 79(3):893–921.

Gradwohl R (2018) Voting in the limelight. *Econom. Theory* 66(1):65–103.

Greiner B (2004) Subject pool recruitment procedures: Organizing experiments with *ORSEE. J. Consumer Res*. 1(1):114–125.

Hansen S, McMahon M, Prat A (2017) Transparency and deliberation within the FOMC: A computational linguistics approach. *Quart. J. Econom.* 133(2):801–870.

Harrison GW, Lau MI, Rutström EE (2009) Risk attitudes, randomization to treatment, and self-selection into experiments. *J. Econom. Behav. Organ.* 70(3):498–507.

Hermalin BE, Weisbach MS (2017) Assessing managerial ability: Implications for corporate governance. Hermalin BE, Weisbach MS, eds. *The Handbook of the Economics of Corporate Governance*, vol. 1 (Elsevier, Amsterdam), 93–176.

Holmström B (1999) Managerial incentive problems: A dynamic perspective. *Rev. Econom. Stud.* 66(1):169–182.

Hossain T, Okui R (2013) The binarized scoring rule. *Rev. Econom. Stud.* 80(3):984–1001.

Irlenbusch B, Sliwka D (2006) Career concerns in a simple experimental labour market. *Eur. Econom. Rev.* 50(1):147–170.

Karamychev V, Visser B (2017) Optimal signaling with cheap talk and money burning. *Internat. J. Game Theory* 46(3):813–850.

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

Keane MP, Runkle DE (1990) Testing the rationality of price forecasts: New evidence from panel data. *Amer. Econom. Rev.* 80(4):714–735.

Keane MP, Runkle DE (1998) Are financial analysts' forecasts of corporate profits rational? *J. Political Econom.* 106(4):768–805.

Koch AK, Morgenstern A, Raab P (2009) Career concerns incentives: An experimental test. *J. Econom. Behav. Organ.* 72(1):571–588.

Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: A tutorial. *Adv. Methods Practices Psych. Sci.* 1(2):259–269.

Levy G (2007) Decision making in committees: Transparency, reputation, and voting rules. *Amer. Econom. Rev.* 97(1):150–168.

Li C, Turmunkh U, Wakker PP (2019) Trust as a decision under ambiguity. *Experiment. Econom.* 22(1):51–75.

Li C, Turmunkh U, Wakker PP (2020) Social and strategic ambiguity vs. betrayal aversion. *Games Econom. Behav.* 123:272–287.

Li Z, Müller J, Wakker PP, Wang TV (2018) The rich domain of ambiguity explored. *Management Sci.* 64(7):3227–3240.

Luenberger DG (2006) *Information Science* (Princeton University Press, Princeton, NJ).

Mattozzi A, Nakaguma MY (2023) Public vs. secret voting in committees. *J. Eur. Econom. Assoc.* 21(3):907–940.

Meade EE, Stasavage D (2008) Publicity of debate and the incentive to dissent: Evidence from the US Federal Reserve. *Econom. J. (London)* 118(528):695–717.

Meloso D, Nunnari S, Ottaviani M (2023) Looking into crystal balls: A laboratory experiment on reputational cheap talk. *Management Sci.* 69(9):5112–5127.

Meyer PE (2014) infotheo: Information-theoretic measures CRAN.R-project. R package version 1.2.0.

Ottaviani M, Sørensen P (2001) Information aggregation in debate: Who should speak first? *J. Public Econom.* 81(3):393–421.

Ottaviani M, Sørensen P (2006a) Professional advice. *J. Econom. Theory* 126(1):120–142.

Ottaviani M, Sørensen P (2006b) The strategy of professional forecasting. *J. Financial Econom.* 81(2):441–466.

Paninski L (2003) Estimation of entropy and mutual information. *Neural Comput.* 15(6):1191–1253.

Prendergast C, Stole L (1996) Impetuous youngster and jaded old-timers: Acquiring a reputation for learning. *J. Political Econom.* 104(6):1105–1134.

Scharfstein DS, Stein JC (1990) Herd behavior and investment. *Amer. Econom. Rev.* 80(3):465–479.

Schlag KH, Van der Weele JJ (2013) Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoret. Econom. Lett.* 3(1):38–42.

Shannon CE (1948) A mathematical theory of communication. *Bell Systems Tech. J.* 27(3):379–423.

Slonim R, Wang C, Garbarino E, Merrett D (2013) Opting-in: Participation bias in economic experiments. *J. Econom. Behav. Organ.* 90:43–70.

Swank OH, Visser B (2013) Is transparency to no avail? *Scandinavian J. Econom.* 115(4):967–994.

Swank J, Swank OH, Visser B (2008) How committees of experts interact with the outside world: Some theory, and evidence from the FOMC. *J. Eur. Econom. Assoc.* 6(2–3):478–486.

Turmunkh U, Van den Assem MJ, Van Dolder D (2019) Malleable lies: Communication and cooperation in a high stakes TV game show. *Management Sci.* 65(10):4795–4812.

Visser B, Swank OH (2007) On committees of experts. *Quart. J. Econom.* 122(1):337–372.