The Strength of Weak Leaders – An Experiment on Social Influence and Social Learning in Teams*

Berno Buechel[1†], Stefan Klößner[2], Martin Lochmüller[3], Heiko Rauhut[4]

[1] University of Fribourg, berno.buechel@unifr.ch

[2] Saarland University, s.kloessner@mx.uni-saarland.de

[3] University of Hamburg, martin.lochmueller@gmail.com

[4] University of Zurich, heiko.rauhut@uzh.ch

April 17, 2019

Abstract

We investigate how the selection process of a leader affects team performance with respect to social learning. We use a laboratory experiment in which an incentivized guessing task is repeated in a star network with the leader at the center. Leader selection is either based on competence, on self-confidence, or made at random. In our setting, teams with random leaders do not underperform. They even outperform teams with leaders selected on self-confidence. Hence, self-confidence can be a dangerous proxy for competence of a leader. We show that it is the declaration of the selection procedure which makes non-random leaders overly influential. To investigate the opinion dynamics, we set up a horse race between several rational and naïve models of social learning. The prevalent conservatism in updating, together with the strong influence of the team leader, imply an information loss since the other team members' knowledge is not sufficiently integrated.

JEL-Code: D83, D85, C91.

Keywords: Social Networks, Confidence, Overconfidence, Bayesian Updating, Naïve Learning, Wisdom of Crowds

*We thank Arun Advani, Sandro Ambuehl, Vincent Buskens, Arun Chandrasekhar, Syngjoo Choi, P.J. Healy, Holger Herz, Matt Jackson, Bernhard Kittel, Michael Kosfeld, Jan Lorenz, Friederike Mengel, Claudia Neri, Muriel Niederle, and Tanya Rosenblat for helpful comments. Berno Buechel gratefully acknowledges the hospitality of the Economics Department of Stanford University and the financial support by the Fritz Thyssen Foundation. Heiko Rauhut acknowledges support by the SNSF Starting Grant BSSGI0_155981.

†University of Fribourg, Department of Economics, Bd. de Pérolles 90, CH-1700 Fribourg, Switzerland; Tel.: +41 26 300 82 55. Email: berno.buechel@unifr.ch. Web: `www.berno.info`.

1 Introduction

In our rapidly changing world, most modern organizations are embedded in highly dynamic environments. For the management of an organization, the first essential step to successful decision-making is the basic task of obtaining an accurate view of the environment.[1] For instance, this can be the foundation for defining a mission statement, as argued, e.g., in Bolton et al. (2013). Recently, there have been a number of contributions showing that organizations can improve their decision-making by harnessing the wisdom of crowds instead of using the expertise of only a single individual (e.g., Surowiecki, 2004; Mannes, 2009; Keuschnigg and Ganser, 2017). However, this literature has not analyzed whether a team's ability to learn from each other depends on characteristics of the team leader.

Given each team member's initial level of information, the updated opinions' accuracy depends on the social learning process within the team. Many teams are organized such that one person, the team leader, directly communicates with each team member, while the other members often communicate only indirectly with each other – via the team leader. In this paper, we address the question of *how the selection of the team leader affects the performance of social learning in the team.* Is it necessary that the central person is the one with the highest expertise? How does self-confidence affect the process of social learning? Should the selection criterion be declared or rather hidden? Answering these questions can be informative for the design of successful organizations.

To address these research questions, we set up a laboratory experiment in which subjects are asked to answer incentivized estimation questions repeatedly. After each round, every team member observes the leader's guesses, while only the leader observes the guesses of all members. We randomly allocate subjects into three treatments, which differ by whether the leader is selected at random, by confidence or by expertise. We use real questions, while previous experiments used highly stylized tasks such as guessing an average (or its sign) of randomly drawn numbers (Çelen et al., 2010; Corazzini et al., 2012) or finding an abstract true state (Choi et al., 2005; Brandts et al., 2015; Chandrasekhar et al., 2015; Grimm and Mengel, 2018). Yet, studying real teams has severe endogeneity problems. For these reasons, we explore the middle ground between theory-testing experiments and field data. We import a method developed outside of economics (Lorenz et al., 2011; Rauhut and Lorenz, 2011), which is increasingly used. Participants are asked to answer knowledge questions about vaguely known facts for which the true answer is known (and could in principle easily be looked up, e.g., on Wikipedia.com). Subjects are paid according to their answers' accuracy and can communicate their confidence levels. The latter aspect is missing in most other experiments of social learning because it is simply not necessary to communicate confidence if signal quality is artificially made common knowledge.[2]

[1] Indeed, disastrous decisions can often be traced back to management teams whose members are in disagreement, or – what is arguably even worse – who unintendedly agree on a distorted view of reality.

[2] Think about the canonical framework with a binary state space and equally precise, conditionally inde-

From a Bayesian perspective, selection of the leader does not matter due to efficient social learning: As it will become clear below, Bayesian learners exchange their opinions such that a consensus is reached independent of who is at the center of the communication network.[3] In contrast, naïve social learning predicts consensus over time with a strong "bias" towards the center's initial opinion.[4] Unless the leader is much better informed than the other team members, this is suboptimal, giving the leader's opinion too much weight. Hence, any leader characteristic that further amplifies the weight of the leader's opinion undermines performance. As such, we study the leader's self-confidence, as well as the public declaration of why the leader was selected.

We assess performance by the proximity of a guess to the correct answer. In particular, we measure the individual and the collective errors of the team's guesses, and use a measure of the wisdom of the crowds. We show that selection of leaders by accuracy or confidence does not outperform random selection. Selection by confidence even undermines performance. Teams with random leaders have the advantage that the non-leaders' guesses are taken into account more strongly when updating information, thereby improving the team's performance. The underlying reason is that declaring the leader as somewhat superior, be it in terms of past performance or past confidence, induces team members to put more weight on the leader's opinion, making the team vulnerable to be misled by a single person.

For a deeper understanding of the opinion dynamics, we further develop rational and behavioral learning models which we compare to our data. Despite a long tradition of theoretical insights and a growing body of empirical research, social learning is still far from being fully understood. Our comparisons between theoretical models and empirical data reveals that people adapt their opinions insufficiently – providing evidence for what is called *conservatism*. While conservatism is common in experiments on belief updating,[5] our extension of social learning models by conservatism is novel. Notice that it is entirely possible that subjects are conservative and at the same time pay too much attention to another subject's opinion. For instance, the declaration of the leader selected by confidence induces our subjects to put too high weights on both, the leader and themselves, at the expense of the weight they can put on the other

pendent signals about the true state. If this is made common knowledge, it is clear how well informed each agent is, and there is no need to communicate confidence. Our technology to provide a confidence level for each estimate is somewhat similar to the literature that considers "tagging" pieces of information with their source (Acemoglu et al., 2014; Phan et al., 2015).

[3]For instance, Gale and Kariv (2003), Rosenberg et al. (2009), and Mueller-Frank (2013) provide frameworks for studying social learning among rational agents who are Bayesian updaters.

[4]For instance, DeGroot (1974), Friedkin and Johnsen (1990), DeMarzo et al. (2003), Golub and Jackson (2010), and Acemoglu et al. (2010) study social learning among naïve agents.

[5]Experiments on belief updating frequently find that real people are more conservative updaters than the theoretical model would predict (Möbius et al., 2011; Mannes and Moore, 2013; Ambuehl and Li, 2018), a pattern that has already been summarized in a classic survey (Peterson and Beach, 1967): "when statistical man and subjects start with the same prior probabilities for two population proportions, subjects revise their probabilities in the same direction but not as much as statistical man does[.]" In this paper, we cannot study the sources of conservative updating, but we can study well the consequences.

group members.

Our paper entails three contributions. First, we provide empirical evidence for advantages of random leader selection (also called sortition, demarchy, allotment, or aleatory democracy). Despite a long tradition of discussion (e.g. Zeitoun et al., 2014; Frey and Osterloh, 2016), empirical evidence is rare and mechanisms are unknown.[6] We demonstrate that declaration of non-random leader selection amplifies the weight of the leader's opinion, which may result in a loss because the wisdom of the crowds in the group is not harnessed. Second, we show that overprecision (or judgemental overconfidence), which is the tendency to provide too narrow confidence intervals for one's estimates (e.g., Soll and Klayman, 2004; Moore and Healy, 2008; Herz et al., 2014) is associated with lower team performance. This suggests that either overprecise leaders should be generally avoided or that the trade-offs between the positive effects of overprecise leaders (e.g., fostering coordination, Bolton et al., 2013; or motivating team members, Gervais and Goldstein, 2007) and their negative impact on social learning should be carefully balanced. Third, our paper makes a methodological contribution. By combining experiments on factual questions with theories on social learning, we build a bridge between neat theoretical frameworks and experimental set-ups that are less stylized. This demonstrates that the assumption of common knowledge about signal precision is problematic. In reality, people do not know the signal precision of their interaction partners, form expectations about it and take into account with which confidence others' opinions are communicated. Moreover, behavioral biases such as overprecision, anchoring effects, or selection bias in information acquisition can give rise to *conservatism* in updating. When incorporating this idea into both naïve and rational models of social learning, we find that each model's fit to the data increases, although the distances to the true answers become larger.

2 Experimental Design

In a nutshell, participants in this experiment were asked to answer the same knowledge questions multiple times in a row. The team leader could observe the previous answers of all team members, while the team members could only observe the previous answer of the team leader. Treatments differed by the selection criterion that determined the team leader.

The experiment was conducted at the University of Hamburg and consisted of eleven sessions with a total of 176 subjects.[7] In each session, participants were randomly allocated into groups of four, which stayed fixed. The basic task was to answer a factual question and to provide

[6] One exception is the study by Haslam et al. (1998), which shows experimentally that randomly selected leaders can enhance team performance in a task of deciding upon priorities in a hypothetical survival situation (e.g., after a plane crash). The mechanism behind the effect, however, remains largely unclear. Interestingly, they also observe that randomly selected leaders are, despite their superior performance, often perceived by their team members as less effective than formally selected leaders.

[7] Participants were mostly undergraduate students from various disciplines; there was no restriction on the pool of participants.

a level of confidence for the answer. The closer the estimate was to the correct answer, the more it was honored by game points which were translated into actual payouts, as detailed in Table C.5 of Online Appendix C.[8] On average, sessions lasted for one hour and participants earned 9.50 Euros, which was close to the remuneration norm of the lab. The maximum feasible payout was 48.20, while the minimum was the show-up fee of 5 Euros. This fact was explicitly stated to the participants in order to highlight that the payout strongly depended on individual performance. It was pointed out verbally and in the written instructions that the use of mobile phones, smart phones, tablets, or similar devices would result in expulsion from the experiment and exclusion from all payments.

Each session consisted of two phases: a selection phase (I) and a treatment phase (II), as illustrated in Table 1. In phase I, each participant answered a set of eight different factual questions. At the end of the experiment, one of these questions was randomly selected to be payoff-relevant. In phase II, there was another set of eight questions, each of which was similar to one of the questions of phase I. For instance, there was a question about voter turnout in both phases of the experiment; similarly, there were two questions about the share of water in certain vegetables. Questions were related to diverse topics and each question had already been tested in previous experiments (Lorenz et al., 2011; Rauhut and Lorenz, 2011; Moussaïd et al., 2013).[9]

In phase II, each question had to be answered six times in a row, i.e., in six consecutive rounds. After each round, participants received feedback about the answers and confidence statements provided by their group members according to a star network, but no other feedback. The center of the star network could observe the previous answers and confidence statements of all four team members; the three pendants could only observe the previous answer and confidence of the center, in addition to their own. For each question of phase II, only one of the six rounds was selected at random by the end of the session to be payoff-relevant. Hence, there was no possibility to "hedge" risk with a portfolio of answers.

The actual treatments differed by the procedure that determined who within a group of four became the center of the star network for phase II. In the baseline treatment T0, the center was selected at random. In the accuracy treatment T1, the group member whose guess on the similar question in phase I was closest to the correct answer was put into the central position of the network. In the confidence treatment T2, this position was given to the group member whose level of confidence for the guess on the similar question in phase I was highest. Potential ties in accuracy or confidence were broken at random. For every question there could be a different center in a given group even when the selection criterion was the same. Half of all groups played the random treatment (T0) for four questions and the accuracy treatment

[8]The chosen payoff function has a convex shape. This provides incentives to report the guess that is most likely the correct answer. Theoretically, an agent's belief is a distribution on an interval and the payoff function is designed to elicit the mode of this distribution, as we explain in subsection B.4.1 of Online Appendix B.

[9]The full list of questions can be found as Table C.1 in Online Appendix C.

(T1) for the other four questions; the other half played the random treatment (T0) for four questions and the confidence treatment (T2) for four questions.[10] The selection procedure was made transparent to the group members when the network for one question was formed, i.e., before the question was answered six times. During phase I, subjects did not know how decisions in phase I could have an influence on phase II. Instructions for the first phase simply announced that there would be a second phase with another set of instructions. This precluded strategic behavior in phase I, e.g., to become the leader or to avoid becoming the leader in phase II. While the answers to the questions were strongly incentivized, the confidence statements were not directly incentivized. Hence, the statements of confidence in phase II can also be considered as a mere communication technology.[11]

Table 1 gives an overview by showing the timeline and the number of observations. First, in phase I, each group was confronted with eight questions in random order. Then, in phase II, it was confronted with the eight corresponding questions in the same order. For the first four questions in phase II, the group was in one treatment, for the latter four in another treatment. In total, this yields 352 unique group-question pairs, of which 176 are in the random treatment T0, 88 in the accuracy treatment T1, and 88 in the confidence treatment T2. Since one group-question pair consists of four people who answer six times the same question (in phase II), our total number of single answers is 8,448.

Sequence	Phase I	Phase II
S1	8Q	4Q under T0, 4Q under T1
S2	8Q	4Q under T0, 4Q under T2
S3	8Q	4Q under T1, 4Q under T0
S4	8Q	4Q under T2, 4Q under T0
Rounds	1	6

Treatment	Groups x Questions
T0 Random	44 x 4 = 176
T1 Accuracy	22 x 4 = 88
T2 Confidence	22 x 4 = 88
Sum	352

Table 1: Overview of the timeline and number of observations. Each of the 44 groups played one sequence. In a sequence, a group answered 8 questions once in phase I and 8 partner questions six times in a row in phase II. This yields 176 group-question pairs in the random treatment T0 and 88 group-question pairs in each of the two other treatments (T1 and T2).

Note that the number of observations in the random treatment T0 was chosen larger in order to have a sufficient number of cases in which by chance the center happened to be the most accurate or the most confident. These cases enable us to disentangle effects of leader selection from effects of declaring of how the leader was selected.[12]

[10]The full schedule of which group played which question in which treatment is given by Table C.3 in Online Appendix C.

[11]As we discuss in the next section, among rational agents there are indeed incentives to communicate truthfully the level of confidence in our setting in order to foster optimal learning in the group. However, our experimental results will not rely on the assumption that the confidence statements are truthful.

[12]A more detailed description of the experimental procedures can be found in Online Appendix C.

3 Theoretical Background

In this section, we derive theoretical predictions about the behavior in our experiment. The set-up is as follows. Let $N = \{1, 2, 3, 4\}$ be the agents in one team. Let 1 be the center of the star network and $2, 3, 4$ the pendants. The basic task in our experiment is to provide guesses on a specific question, the answer of which is a fraction. Hence, there is an unknown state of the world $\theta \in \Theta$, which is the correct answer to the question at hand.[13] Denote by $x_i(t)$ the answer of agent i at time t. Denote by $c_i(t)$ the confidence statement of agent i at time t. Time is discrete: $t = 1, 2, ..., T$, with $T = 6$ in phase II of the experiment. Accurate guesses are incentivized by a payoff function $\pi(e_i(t))$ that is weakly decreasing in the distance to the true answer, $e_i(t) := |\theta - x_i(t)|$. One out of six answers is finally drawn as payoff-relevant.

To make predictions about the participants' guesses in phase II, we use two approaches: a rational learning approach and a naïve learning approach.

3.1 Rational Learning Approach: Bayesian Updating

In the rational learning approach, we assume that agents maximize expected payoffs given their beliefs and that beliefs are formed by Bayes rule.

Notice that a belief about the true answer is not a single number, but a probability distribution over the possible states ($f_i(t) : \Theta \to \mathbb{R}$). In the first round of guessing, $t = 1$, agents are endowed with some private information, i.e., what they know about the question at hand before interacting in the team. In the second round, each pendant $i \neq 1$ has observed the guess $x_1(1)$ and the confidence statement $c_1(1)$ of the center and can use this to update his belief. The center, on the other hand, has observed all guesses and confidence levels of the first round to form her belief, which is the basis for her second-round guess $x_1(2)$.[14] If we assume that the guess and confidence level are sufficient to reconstruct an agent's belief and that the agents know how their private information is interrelated, then the center is fully informed after the first round of guesses. In this case, she can make the optimal guess $x^* := \arg\max_{x \in \Theta} E[\pi(|\theta - x|)|f_1(1), ..., f_4(1)]$, given the pieces of information in the team. Since all agents have the same payoff function and pendants can observe the center's guess $x_1(2) = x^*$, all agents make the same guess $x_i(t) = x^*$ from round 3 on. This observation leads to the following prediction.[15]

Prediction 1 (Bayes)**.** *In a model with common knowledge of rationality and common priors, the following holds. If the answer and confidence statement of a linked team member in a star*

[13] In the experiment, the correct answer is rounded and belongs to the finite set $\Theta = \{0, 0.01, 0, 02, ..., 0.99, 1\}$, which we can also model as the interval $\Theta = [0, 1]$.

[14] For easier readability, we use the female form for the center and the male form for the pendants.

[15] A formal statement of this result can be found in Online Appendix B. There we introduce the general framework (B.1), prove the proposition (B.2), and provide two specific examples how such a rational model unfolds in our setting (B.4.1).

network is sufficient to fully represent her private information, then the center learns once and the pendants learn twice. (Learning refers here to information updates and improvements in expectations.) Moreover, all team members will state the optimal answer x^* in any round $t \geq 3$, independent of who is at the center of the star network.

Prediction 1 states that the selection of the team leader does not matter for the performance of social learning, apart from the first two rounds (and, in fact, only apart from round two). Moreover, it states that every agent provides the payoff-maximizing guess, which implies that social learning is "efficient" in the sense of maximizing the sum of expected payoffs.[16] However, several of its underlying assumptions deserve further attention.

First, it is explicitly assumed that statements of guesses and confidence levels are sufficient to recover beliefs. For this to be satisfied, the agent must know the other's belief up to one or two parameters. This is satisfied, for instance, in models assuming that beliefs follow a beta distribution.[17] Bayesian models with weaker assumptions could assume that agents also have beliefs about the signal quality of the others and imperfectly learn over time both the available private signals as well as their quality. Given the result by Aumann (1976), such a model is expected to lead to more learning iterations, but to the same outcome in the long run.

Second, how exactly an agent updates depends on his higher order beliefs on how private pieces of information are related to each other and how they are related to the truth. In theoretical models, it is usually assumed that there is common knowledge about the prior distribution of the true state, and about how private signals are drawn. In this experiment, agents are confronted with real questions. Hence, the agents' higher order beliefs about their own and their fellow team members' expertise can also depend on additional factors, such as the particular question at hand or on the treatment. In particular, the accuracy treatment T1, i.e., that the center gave the most accurate answer to a similar question, or the confidence treatment T2, i.e., that the center was the most confident on a similar question, might reveal something about the agent's ability that could be considered in the updating process. If anything, the declaration of the treatment T1 or T2 can reveal additional information, which would lead to better guesses, compared to the random treatment T0. To generate a prediction that is much more in line with the theoretical models, Prediction 1 abstracts from this possibility by assuming that there is common knowledge about how the private pieces of information are related to each other and to the truth.[18]

Third and finally, the assumption of common knowledge of rationality need not be satisfied. In sum, it cannot be expected that the requirements of Prediction 1 above are fully satisfied in the experiment. Still, Prediction 1 gives us a clean baseline to compare the data to.

[16] Since efficiency here means that not only the sum but also each individual's expected payoffs are maximal, there are no incentives to deviate, e.g., by misrepresenting the own opinion or confidence level.

[17] We study such models in section 5. They are formally introduced in Online Appendix B.4.

[18] In the experiment, we did not induce a common prior because we used questions of real topics. Nevertheless, we argue that models that assume a common prior and signals can contribute to our understanding of social learning in real settings.

3.2 Naïve Learning Approach: DeGroot Model

Previous experimental research on social learning has not always found strong support for Bayesian learning, but often suggests that simple rules of updating, such as repeatedly taking averages, fit the data well (Corazzini et al., 2012; Battiston and Stanca, 2015; Chandrasekhar et al., 2015; Grimm and Mengel, 2018). We use their common modeling approach, which is often named after Morris DeGroot, to generate an alternative prediction and to later specify models of more naïve learning. The basic aspect of naïveté incorporated in this modeling approach is that agents do not sufficiently account for the origin of information such that pieces of information are used each time they reach an agent through the network. This behavioral bias is also called "persuasion bias" (DeMarzo et al., 2003).

In the DeGroot model, the way people average the former guesses in their network neighborhood is typically constant. In the star network, this means that peripheral agents always provide a guess that is a mixture between the center's and their own last guess, with constant weights g_{i1} and g_{ii} on the two, while the center mixes all answers with some constant weights $g_{11}, g_{12}, g_{13}, g_{14}$, which are also positive and sum up to one. Given the weights and the initial answers $x_i(1)$, all consecutive answers $x_i(t)$ are fully determined. In particular, if G denotes the (row-stochastic) 4×4 matrix consisting of these entries g_{ij} and zeros at the remaining entries, the agents' updating can be written in vector and matrix notation as $x(t) = Gx(t-1)$. Hence, the predicted guesses are $x(t) = G^{t-1}x(1)$, for $t = 1, 2, ...$ Each agent thus generically changes guesses from round to round. Assuming that averaging weights are strictly positive is sufficient for the conclusion that all agent's guesses $x_i(t)$ converge for $t \to \infty$ to the same answer, which we denote by $x_i(\infty)$. Given that convergence is fast enough, $x_i(\infty)$ is also a good prediction for $x_i(6)$. For the star network, it can be shown that, for any i,

$$x_i(\infty) = \frac{1}{c}\left(1x_1(1) + \frac{g_{12}}{g_{21}}x_2(1) + \frac{g_{13}}{g_{31}}x_3(1) + \frac{g_{14}}{g_{41}}x_4(1)\right), \tag{1}$$

with $c = 1 + \frac{g_{12}}{g_{21}} + \frac{g_{13}}{g_{31}} + \frac{g_{14}}{g_{41}}$. The weights $w_i = \frac{1}{c} \cdot \frac{g_{1i}}{g_{i1}}$ measure long-term influence of an agent i, which is called eigenvector centrality in network science since $w'G = w'$ (e.g. Friedkin, 1991; DeMarzo et al., 2003; Golub and Jackson, 2010). As can be directly observed from Equation (1), the center's influence on the long-term answer is different from a pendant i's influence, as long as $\frac{g_{1i}}{g_{i1}} \neq 1$. In particular, the center has a stronger influence if the center's weight on the pendant g_{1i} is lower than the pendant's weight on the center g_{i1}. This is a realistic assumption since pendants have only the center's guess to update from, while the center can distribute her weight among three pendants.

To discuss performance of social learning in this model type, we need to make assumptions about the relation between the initial guesses $x_i(1)$ and the truth θ, e.g., that initial guesses are realizations of independent random variables that have the truth as expected values. For any such probabilistic model and for any definition of the "optimal" guess $\hat{x}$ given the initial

guesses, the approached value $x(\infty)$ and the optimal guess $\hat{x}$ will only coincide if by coincidence the averaging weights happen to be optimal in that sense. The same holds true for the guesses and optimal guesses of early rounds, say round two. Even if the weights g_{ij} happen to produce the optimal guess $\hat{x}$ for some agent i in some round t, they will not have this property for every agent and for every round. Hence, there is an inherent inefficiency in these naïve models of social learning. The reason is that initial guesses of some participants are incorporated in the change of answers more frequently than other team members' guesses, while guessing weights are constant. These observations lead to the following prediction.[19]

Prediction 2 (DeGroot)**.** *In the naïve model with constant and positive averaging weights, the following holds. In a star network, every agent's learning heavily depends on the network positions, i.e., on who is the center. In particular, for $g_{i1} > g_{1i}$, the center has a larger influence on the long-run opinion than team member i. Generically, the center updates more than once and the pendants update more than twice. Under weak conditions, the first round of updating is learning (the expected error decreases), but for every notion of what is the optimal answer, all team members will generally state suboptimal answers.*

Prediction 2 states that the selection of the team leader heavily affects the performance of social learning, and that social learning is generally "inefficient" in the sense of not maximizing any function that is decreasing in the error of an agent's guess. Given the weighting matrix G, the naïve model is fully specified and provides a clear-cut prediction about all agents' guesses in all rounds. Typical specifications of G are studied in section 5.2.

Our treatments T1 and T2 mainly affect naïve social learning through the manipulation of the network positions (who is at the center), but potentially also through the declaration of the treatments. The second channel would be present if the averaging weights g_{ij} depended on this declaration. In the empirical analysis, we will disentangle the effects of the manipulation of the center – which does not matter according to Prediction 1, but is crucial according to Prediction 2 – from potential effects of declaration (which can only be helpful in the rational framework of Prediction 1, but could also be harmful in the naïve framework of Prediction 2).

4 Success of Social Learning

The two theoretical approaches lead to contradicting predictions. Therefore, it remains an empirical question whether and how the selection of the leader affects the success of social learning.

[19] A formal statement of this result can be found in Online Appendix B. There we introduce a probabilistic framework and prove the proposition (B.3); and also provide two specific examples (B.5.1).

4.1 Performance over Time

We measure performance both on the individual and on the collective level. We define the individual error $e_i(t)$ by the absolute distance between answer $x_i(t)$ and truth θ. On the group level, we use two complementary measures. We define the *collective error* by the error of the mean of the four answers in the group $ce(t) = |\frac{1}{4}\sum_{i=1}^{4} x_i(t) - \theta|$. We define the *(wisdom of) crowd error* by the degree as to whether the answers "bracket" the true value (following the spirit of Lorenz et al., 2011). Accordingly, we define $woce(t) = 0$ if at most two answers are strictly below or strictly above the correct answer; $woce(t) = 1$ if three answers are strictly below or strictly above the correct answer; and $woce(t) = 2$ if the correct answer lies strictly above or below all four answers in the group.[20]

Figure 1 depicts the levels of these performance measures over time by the three treatments. Panels A-C show that the individual errors are on average between 10 and 20 percentage points from the true answer and decrease over time. As intended, in the accuracy treatment T1, selecting a center who was most accurate in answering a similar question (in phase I) leads to centers who are significantly better in estimating the current question in the first round (of phase II), while this is, notably, not the case in the confidence treatment T2. The centers' individual errors reduce significantly in the random and confidence treatment, but not in the accuracy treatment. By and large, this is consistent with rational learning models (which take guess and confidence as a sufficient statistic for someone's belief), i.e., that pendants learn twice and centers once.[21] Panels D-F show that collective errors also first decrease and then settle.[22] Taking these observations together, agents mostly learn in the first and second round of updating.[23] A similar pattern, albeit with a necessary change of sign, can be observed in panels G-I for the crowd error: The crowd error increases over time with most of its changes until round $t = 3$. Hence, in the final period the correct answer most frequently lies outside the "bracket" of all provided answers. This observation is consistent with findings of Lorenz et al. (2011).

Result 1. *Individual and collective errors reduce over time. Centers learn once (except in the accuracy treatment T1); pendants learn at least twice. Crowd errors increase over time.*

[20] Thus, the crowd error measures whether the correct answer lies within the interval that is spanned by the four answers, and if so, whether it also lies within the interval that is spanned by the two answers which are contained in the interval of the two other answers. "Bracketing" is important when the decision maker assumes that the truth lies in the interval spanned by the answers.

[21] Recall that we derived the predictions from the Bayesian approach using the assumption that guess and confidence taken together are a sufficient statistic for someone's belief. If this assumption fails, higher order beliefs matter and more rounds of learning are expected.

[22] The apparent differences between treatments in the first round of the collective error are neither significant, nor are they driving the subsequent results, as it can be shown.

[23] Learning cannot stem from having more time to think about a question since subjects who are not confronted with any information about the guesses and confidence of others did not at all improve over time. We tested this possibility with subjects who were randomly drawn from all potential participants in sessions whose number of potential participants was not divisible by four, the size of our groups.

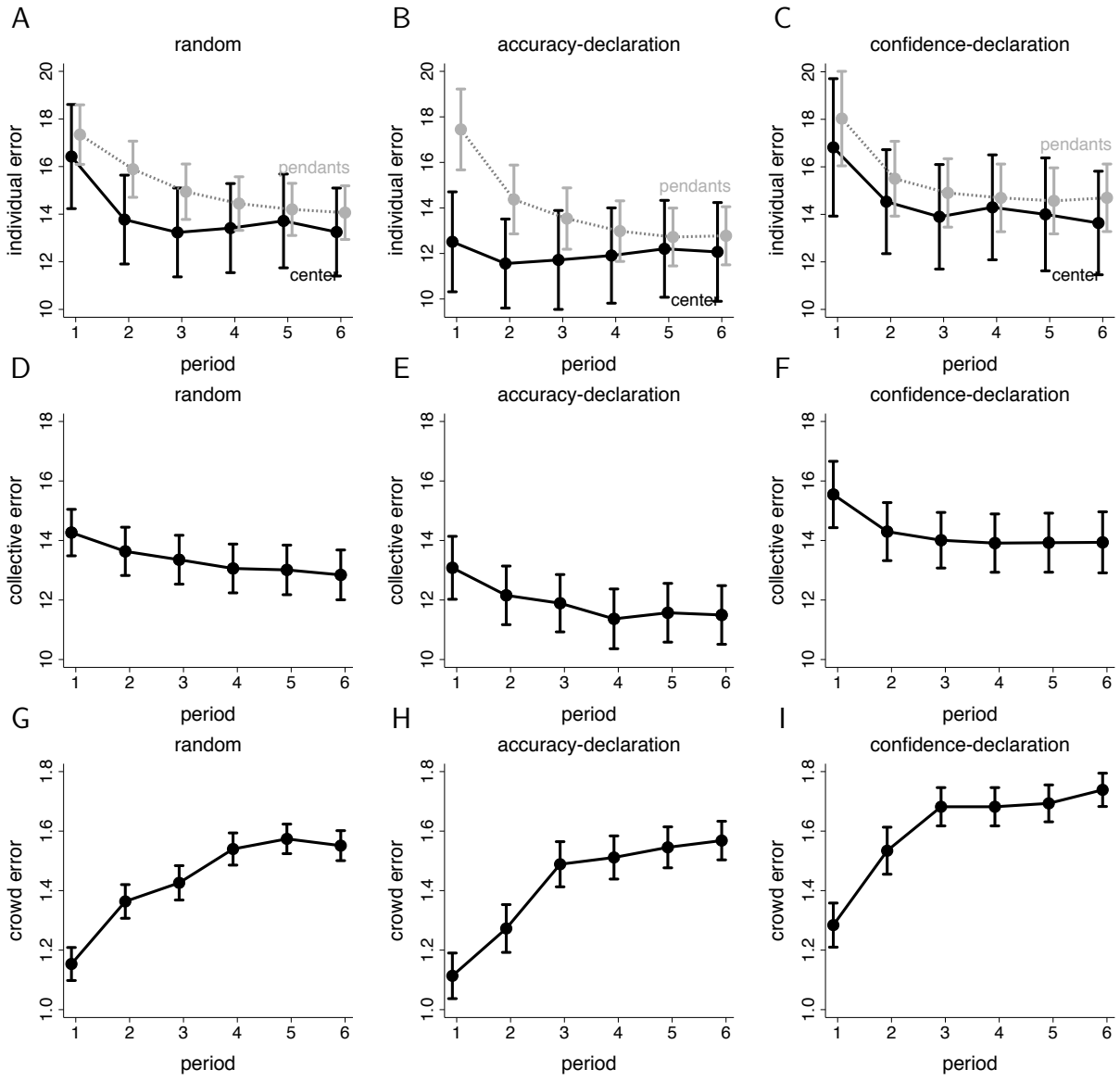

Figure 1: Individual, collective, and crowd errors over time by treatments. Panels A, B, C differentiate between centers (black) and pendants (gray). All confidence intervals are standard 95% confidence intervals.

Another view on the change of error over time is provided by Figure A.1 in the Appendix. It shows for each question the distribution of the first round and the last round answers, indicating a substantial heterogeneity between questions, for which we control in the subsequent analysis.

4.2 Treatment Effects on Performance

To test for treatment effects, we run regressions with the three error measures as the dependent variables and with treatment dummies as the independent variables. We focus our analysis on investigating the effects of learning on the final period, which is period 6. The last period is the most relevant, since it is the last period up to which learning can take place. In consecutive robustness analyses, we also analyze performance for earlier rounds back to period $t = 3$, the first round in which full learning can theoretically take place. Notice that the distribution of (individual and collective) errors is heavily skewed. Taking the logarithm (e.g., $\log(e_i(t)+1)$) in the regressions of individual and collective errors gives less weight to errors which are far away from the truth and more weight to errors close to the true answer, such that the analysis will not be driven by a few cases in which errors were huge, say, forty and more. For the variable crowd error, which may attain values 0, 1, and 2, we use ordered logit.

	(1)	(2)	(3)
	individual error (log)	collective error (log)	crowd error
accuracy treatment (T1)	0.026	0.003	0.106
	(0.30)	(0.03)	(0.39)
confidence treatment (T2)	0.144	0.179	0.739*
	(1.58)	(1.78)	(2.38)
intercept	2.164***	2.149***	
	(22.65)	(18.07)	
intercept cut 1			-2.555***
			(-6.85)
intercept cut 2			-0.830*
			(-2.44)
N	1'408	352	352

Question dummy coefficients for 8 questions not shown

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Treatment effects on final errors: log individual error, log collective error, and wisdom of crowd error (in period 6). Linear regression (models 1 and 2) and ordered logit regression (model 3). The reference category is the random treatment T0.

Table 2 reports these models when controlling for each treatment T1 and T2 with a dummy variable, while T0 is the reference category. We control for possible heterogeneity between different questions by using corresponding dummy variables. Throughout all analyses, we use robust standard errors. They are clustered for the combination of group and question to account for inter-dependencies within a group when answering the same question. If selecting the most accurate or the most confident enhances performance, then we should see a significant

negative effect on the three errors. As Table 2 reveals, the accuracy treatment T1 and the confidence treatment T2 do not outperform the random treatment T0. The coefficients are mostly insignificant and in fact positive. There is even some indication that the confidence treatment T2 underperforms compared with the random treatment T0. The latter effect is significant at the 5% level for the crowd error, while the null hypothesis cannot be rejected for collective error ($p = 0.075$) and individual error ($p = 0.114$).[24] To further investigate the potential negative effect of the confidence treatment T2 on the individual error, we rerun the regression with the expected payoff in EUR as the dependent variable (see model (1) of Table A.1 in the Appendix). It turns out that the effect is significantly negative ($p < 5\%$) and can be quantified as follows: Being in T2 in comparison to T0 reduces the expected utility for the last round guess for every question by 0.17 EUR. This is a decrease of 36% from the reference value 0.48 (see intercept of model 1 in Table A.1).

Result 2. *Performance does not improve when the center is known to be the most accurate (T1). Performance deteriorates when the center is known to be the most confident (T2).*

To understand the mechanism behind these treatment effects of selecting the most accurate or the most confident agent as a center, we distinguish between two aspects of each treatment, the trait of the central agent and the declaration of how the central agent was selected.[25] By our experimental design we can disentangle the two effects, since in the random treatment T0 it frequently happens by chance that the most accurate agent was selected as the center without having the declaration of her accuracy, as is the case in the T1 treatment. The same applies for confidence; in a number of cases, the most confident agent was randomly selected to be the center in the random treatment T0.

Table 3 reports the results of the regressions when we control for the trait that the center is the most accurate or the most confident in the group, such that the treatment dummies only pick up the declaration effect. When the center happens to be the most confident or the most accurate (in the corresponding question of phase I), the outcome measures tend to improve, which is indicated by the negative sign of the (mostly non-significant) coefficients. When the confidence of the center is declared to all group members, however, the performance is significantly reduced. To quantify this effect, we rerun this regression using again the expected payoff in EUR as the dependent variable (see model (2) of Table A.1 in the Appendix). Declaring that the center was the most confident (T2), the expected payoff reduces by 0.26 EUR. This is a decrease by 49% from 0.54 EUR in the case of having the most confident in the center in the random treatment for the baseline question. The results are qualitatively similar for accuracy of the center in the sense that the signs of the coefficients are the same, but we cannot reject the null in that case, and the size of the effects is also smaller than for confidence.

[24] In the regression tables we report the t-statistics, which can be transformed into the p-values. The tests are two-sided.

[25] For easier readability, we often only write the most confident or the most accurate center without explicitly repeating that this refers to confidence and accuracy in the corresponding question of phase I.

	(1) individual error (log)	(2) collective error (log)	(3) crowd error
accuracy-trait	-0.110	-0.0716	-0.0477
	(-1.13)	(-0.69)	(-0.15)
accuracy-declaration (T1)	0.117	0.0790	0.196
	(1.01)	(0.64)	(0.53)
confidence-trait	-0.106	-0.231*	-0.474
	(-1.19)	(-2.21)	(-1.74)
confidence-declaration (T2)	0.218*	0.335**	1.053**
	(1.98)	(2.66)	(2.90)
intercept	2.221***	2.241***	
	(22.42)	(18.81)	
intercept cut 1			-2.735***
			(-7.31)
intercept cut 2			-0.999**
			(-2.92)
N	1'408	352	352

Question dummy coefficients for 8 questions not shown

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Treatment effects on final errors: log individual error, log collective error, and wisdom of crowd error (in period 6). Linear regression (models 1 and 2) and ordered logit (model 3). The reference category is the random treatment T0 restricted to the cases where the center has neither the accuracy-trait nor the confidence-trait.

While Table 3 reports the effects for the final period after all learning has taken place, Figure 2 illustrates robustness analyses of declaration effects when the regressions are run for each period separately. We show periods 3 to 6, since these are the periods after which full learning could happen and did take place according to the error dynamics (Figure 1).

The effect of declaring that the center is the most confident consistently increases the error measures and thus reduces performance. The declaration of accuracy has the same tendency, but the effects are smaller and insignificant.

Result 3. *Declaration of confidence undermines performance.*

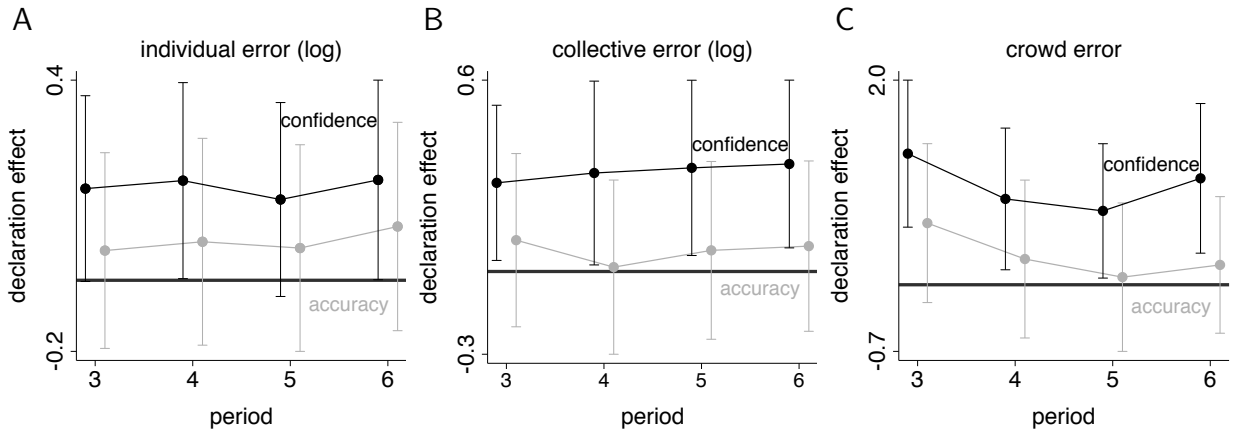

Figure 2: Treatment effects on errors: log individual error, log collective error, and wisdom of crowd error (periods 3-6). Linear regressions, 95 % confidence intervals.

4.3 Social Influence

To analyze why the selection of the center may have a negative impact on performance, we study to which extent agents within a group influence each other. For this purpose, we regress the answer $x_i(t)$ of an agent i at time $t \geq 3$ on his initial answer $x_i(1)$, as well as on the initial answers of the other group members $x_j(1)$. In particular, a pendant's answer is regressed on the center's initial answer, his own initial answer, and the mean of the other two pendants' initial answers. The center's answer is regressed on the average of the pendants' initial answers.

Tables A.2 and A.3 in the Appendix report the influence weights on $t = 6$ when estimating them separately for each treatment. For instance, in the random treatment T0, a pendant's final answer is estimated as the convex combination of its initial answer with weight 56.7%, the center's initial answer with weight 26.7%, and the other pendants' average initial answer with weight 16.6%. There are several interesting observations contained in these tables. First, every agent places much weight on his own initial opinion. In the rational model and the random treatment, we would expect that on average this weight is 25%.[26] Second, the weight individuals place on their own initial opinion depends on the treatment. In the random treatment T0, pendants place more weight on themselves, while centers place less weight on themselves. Finally, the social influence by the other team members heavily depends on the treatment. For pendants, the center's weight was 26.7% in the random treatment T0, but 46.9% in the confidence treatment T2; and similarly in the accuracy treatment T1.

The two aspects of a treatment, the trait of the center and the declaration of how the center was selected, are then captured by the interaction effects of the corresponding dummy variables with the influence weights in the regressions that pool the three treatments. These regressions are reported in Tables A.4 and A.5 in the Appendix, their effects are illustrated in Figure 3.

[26]We will return to this observation when extending the social learning models in section 5.

A positive effect of a certain dummy variable thereby means that the corresponding influence weight is being increased by the corresponding treatment.

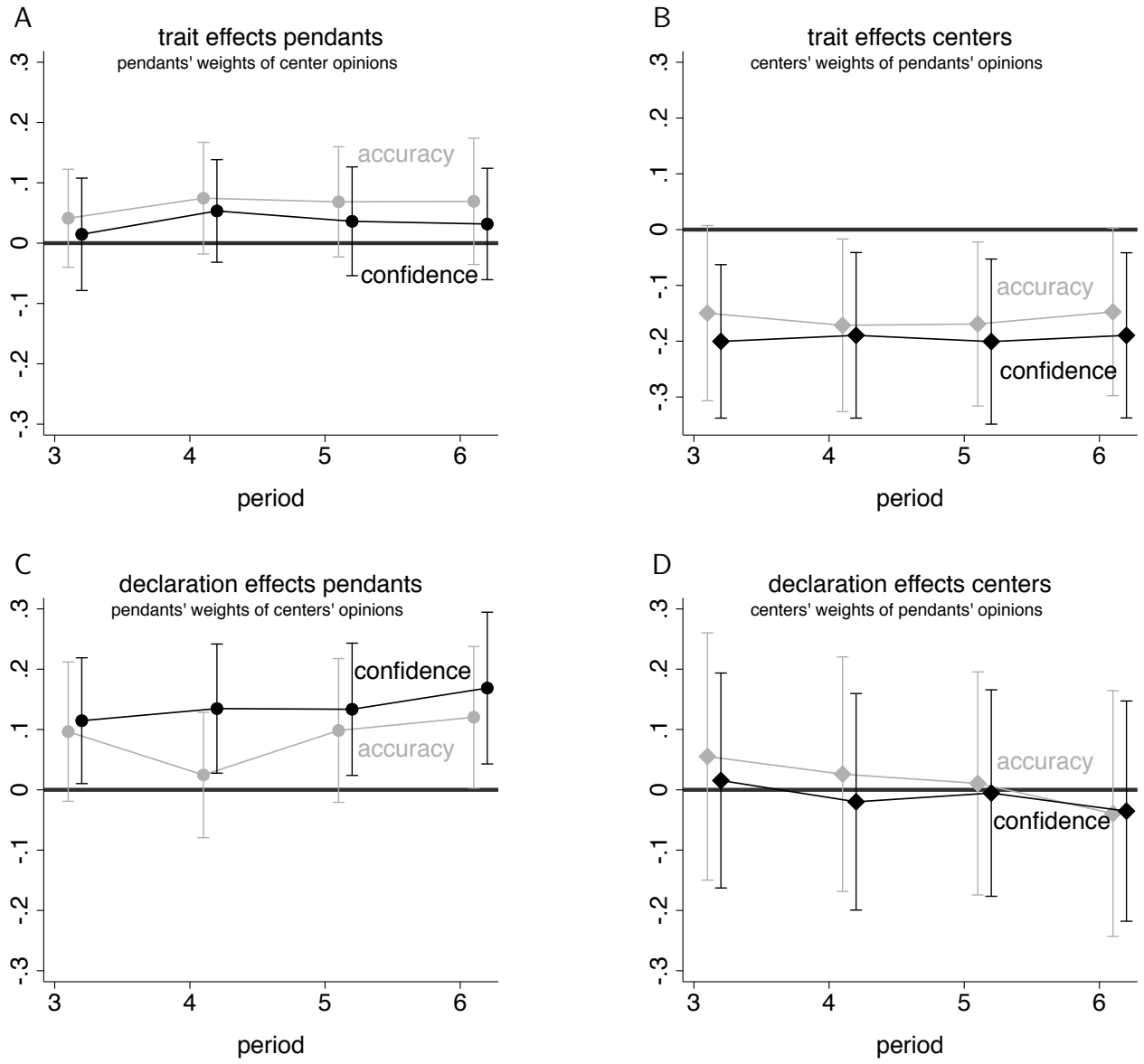

Figure 3: Trait and declaration influence for pendants and centers. Gray accuracy, black confidence treatments, 95 % confidence intervals. Panels A and C show how a pendant's answer in late periods is influenced by the center's initial answer. Panels B and D show how a center's answer in late periods is influenced by the pendants' initial answers.

When the center happens to be the most accurate or the most confident in phase I, but there is no public declaration of this, then the pendants do not strongly respond (panel A), while the center places significantly more weight on her own initial opinion and, accordingly, significantly

less weight on the pendants' opinions (panel B). In contrast, the declaration that the center is the most confident or accurate does not affect the center's weighting (panel D), but there is a strong effect on the pendants (panel C). Declaring that the center is somehow special (the most confident or accurate on a similar question) significantly increases the pendants' weights on the center's initial opinions.

Result 4. *The pendants place more weight on a center who is* declared *to be the most confident or most accurate. The center places less weight on the pendants when she* is *the most confident or the most accurate.*

This result provides an explanation for the former results. Intuitively, placing more weight to a single opinion has a negative effect on performance, except if this person is substantially better informed than the others. In the accuracy treatment T1, this condition is satisfied to some extent, such that the negative effect of placing too much weight on a single person and the positive effect of placing more weight on a person who is better informed may balance each other. Consequently, the performance in the accuracy treatment T1 need not differ from the random treatment T0. In the case of the confidence treatment T2, the center is not substantially better informed than the other group members, as can be seen from panel C in Figure 1. Hence, putting more weight on her initial guesses only has the negative effect of insufficiently taking into account the information of the others. This may lead to performing worse than under the random treatment T0.

4.4 Overprecision

It is well-known that many people often suffer from a form of overconfidence called overprecision, i.e., they report much too small confidence intervals when asked about a region where they expect the true answer with a certain probability (a usual way is to ask where they expect the answer in 90% of their guesses; see, e.g., Soll and Klayman, 2004; Moore and Healy, 2008; Herz et al., 2014). In phase I of our experiment, we asked participants to provide such regions. Therefore, we can compute for every participant the individual overprecision score simply by counting how often that person provided a confidence interval that did not contain the true answer. Thus, every participant is characterized by an overprecision score in $\{0, 1, \ldots, 8\}$. As Figure 4 reveals, many agents are overprecise. Their guess should only lie in 10% of the cases outside of their provided 90% confidence interval. However, for most agents this happens in more than two out of eight cases. The histogram also documents that there is substantial heterogeneity in overprecision.

In model (3) of Table A.1 and Table A.6 in the Appendix, we analyze how the center's overprecision score as well as the average of the pendants' overprecision scores impact the group's performance (on top of the previously found treatment effects). We first find that the formerly discussed effects (in particular, the negative declaration effect of T2) remain significant

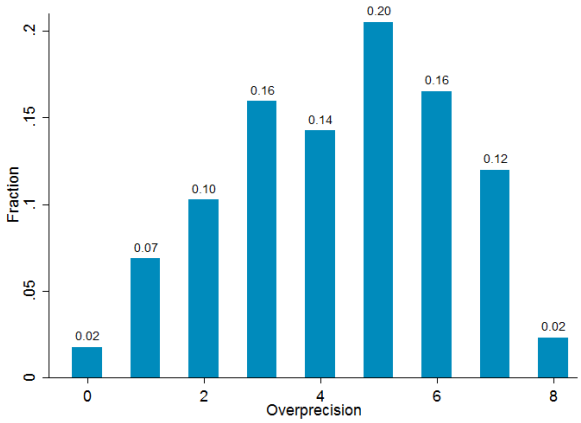

Figure 4: Histogram of individual overprecision. The value 0 means that a subject has specified for all eight knowledge questions a respective 90% confidence interval which encloses the true value. The value 8 means that a subject has specified for all eight knowledge questions a 90% confidence interval which does not contain the true value. All values above 1 indicate overprecision, since more than 10% of estimates fall out of the 90% confidence interval (i.e., 91% of subjects are overprecise).

and are hence robust when controlling for overprecision. Second, we observe that the center's and the pendants' overprecision coefficients are positive in Table A.6 and negative in Table A.1. They are significant when the dependent variable is the expected payoff (Table A.1, model (3)) or the crowd error (Table A.6, model (3)). For the individual and the collective error, these effects are not significant on the 5% level, with p-values that are all between 5% and 10% (as reported in Table A.6). Taken together, we interpret this as sufficient evidence for the following result.

Result 5. *Both the center's and the pendants' overprecision are associated with lower performance.*

Tables A.1 and A.6 additionally indicate that the center's overprecision score has a more deteriorating impact on performance than a pendant's overprecision score. Therefore, ceteris paribus, it is best for the group's performance if the least overprecise group member was the center. On the other hand, overprecision is related to confidence, and the group member most confident in phase I acting as center might improve the group's performance when she is not declared to be the most confident. Indeed, Table A.6 shows that, when controlling for overprecision and for the declaration of confidence, the trait of being the most confident significantly increases performance with respect to the collective error and the crowd error. However, this effect is not significant for the individual error (model (1) in Table A.6, $p = 0.140$) and the expected payoff (model (3) in Table A.1, $p = 0.055$). Thus, we conclude that the leader personality that should optimally be selected may well be characterized as confident, but not

as being overprecise. Hence, all results (Results 1-5) contribute to a coherent picture of how the selection of the leader affects social learning.

In the next section, we connect the data more to the theory of social learning. As the social influence analysis showed, both pendants and centers generally placed much weight on their own initial opinion. When studying the learning behavior in the next section, we will incorporate this behavioral aspect and study its consequences.

5 Learning Behavior

5.1 Specification and Extension of Learning Models

We now study how several model variations of both the rational and the naïve model class fit to the data. Table 4 provides an overview of the considered models.[27] The first four models stem from the rational approach to social learning: the *Standard Model* supposes that all agents in a group are equally well informed; the *Sophisticated Model* supposes that signal precision can be derived from an agent's guess and confidence statement. The next four models stem from the naïve approach to social learning: in the *DeMarzo et al. Model* pendants put equal weight on their own and the center's opinion in any round of updating; in the *Corazzini et al. Model* they put higher weight on the center's guesses. Each model has a counter-part, in which conservatism is introduced, indicated by "Plus".[28] For both model classes, introducing conservatism not only reduces the adoption of others' answers in the early rounds, but also alters the prediction that consensus is reached or approached. Conservatism leads to the prediction that the agents' answers are swayed toward their own initial opinion. Taking this idea to the extreme, we obtain the *Sticking Model*, in which every agent sticks to his initial guess without changing it, a simple baseline model.

We implement each model such that all periods $t \geq 2$ are predicted from values at $t = 1$. We can not only assess how well these models fit to the data, but also how close the model predictions are to the true answers.

5.2 Comparison of Models (Horse Race)

We assess the fit of each model by measuring the root of the mean squared error (RMSE) between the model predictions for $t \geq 2$ and the data points, Figure 5 displays the results.

The worst overall model fit is obtained by the baseline model, in which all agents stick to their initial guess (*Sticking Model*). The best model fit is obtained by the "Plus" models, which incorporate conservatism. In fact, every model considered has a larger RMSE than its

[27] The models are formally defined and characterized in sections B.4 and B.5 of Online Appendix B.

[28] In the rational learning models, we derive conservative behavior from the assumption of overprecision (cf. subsection B.4.2 of Online Appendix B). In the naïve learning models, we base conservative behavior on a framework from Friedkin and Johnsen (1990) (cf. subsection B.5.2).

Model	Class	Weighting of others	Conservatism	Consensus
Standard	rational	equal	no	reached
Standard Plus	rational	equal	yes	no
Sophisticated	rational	according to confidence	no	reached
Sophisticated Plus	rational	according to confidence	yes	no
DeMarzo	naïve	equal	no	approached
DeMarzo Plus	naïve	equal	yes	no
Corazzini	naïve	according to degree	no	approached
Corazzini Plus	naïve	according to degree	yes	no
Sticking	both	no	totally	no

Table 4: Overview of model specifications.

"Plus" counterpart that incorporates conservatism. Considering the model fit for each round separately, the conservatism aspect seems particularly helpful in predicting the first updates (round 2), but the effect also persists to the latter rounds.

We can also differentiate model fit by treatment and by the role of being a center or a pendant (see Figures A.2 and A.3 in the Appendix). The most important insight is that the "Plus" models always fit better than their counter-parts. The result holds for all four considered models, for all three treatments, for all rounds, and, apart from one exception, for both centers and pendants.[29] Hence, there is overwhelming evidence for the first part of our Result 6 below.

There are some additional observations to make in Figures A.2 and A.3 in the Appendix. The best model fit in the random treatment T0 is obtained for both the *DeMarzo et al. Plus Model* and the *Standard-Plus Model* with an RMSE of 7.88. Hence, these extensions of straightforward specifications of the naïve and the rational approach best predict the experimental data in the baseline treatment. The *Corazzini et al. Model*, which predicts an immense influence of the center, fits better in the accuracy T1 and confidence treatment T2 than in the random treatment T0 and it fits well for the center, but not for the pendants. The reason is that the center is given a high influence weight in the accuracy and confidence treatment, as well as in the *Corazzini et al. Model* specification. Complementarily, the baseline model of sticking to the initial guess fits much better in the random treatment T0 than in the others. This is a clear indication that social influence is weakest in the random treatment T0 and stronger in the accuracy treatment T1 and the confidence treatment T2. Given that social influence can undermine the wisdom of crowds (Lorenz et al., 2011), this is an explanation for our result that the crowd error is lowest under the random leader T0.

Finally, we can not only assess how these models fit to the data, but also how close the model predictions are to the correct answers. Figure A.4 in the Appendix displays how far the answers based on these models lie from the truth. This overview indicates that the introduction

[29]The exception is that the *Corazzini et al. Model* predicts the center's guesses better than the *Corazzini et al. Plus Model*. Recall that the center already has a high weight on herself in the *Corazzini et al. Model* model.

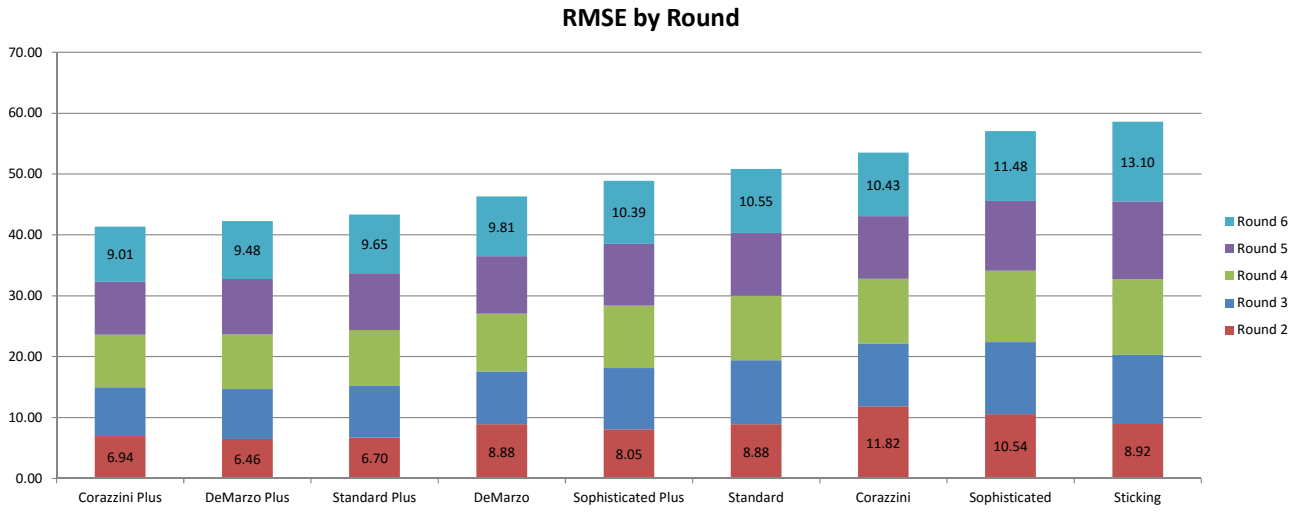

Figure 5: Root mean squared errors (RMSE) of social learning models. "Standard" and "Sophisticated" are models of rational learning; "DeMarzo" and "Corazzini" are models of naïve learning. "Plus" models incorporate conservatism. Lower errors mean better fit between model and data.

of conservatism does not improve the guesses, since the "Plus" models are further from the truth than their counterparts. The same observation holds for all four considered models, for all three treatments, for all rounds, and, apart from two exceptions, for both centers and pendants.[30] Hence, there are two main findings from the horse race as summarized by the following result.

Result 6. *Incorporating "conservatism" into both the rational and naïve models of social learning increases the fit between theoretical models and empirical data. It, however, decreases the fit between the theoretical models and the correct answer.*

The first statement strongly indicates that the extension of both the rational and the naïve models of social learning by conservatism is not a mere theoretical exercise, but an empirically relevant generalization. The second statement shows that conservatism is usually harmful. However, it must be noted for this latter statement that conservatism has different effects on different measures of performance. For instance, the "Plus" models perform better in terms of the crowd error than their counterparts.[31]

[30] These manifold comparisons are not all reported in the paper. The exceptions are the centers in the rational models (*Standard Model* and *Sophisticated Model*) who are left better off when no group member is conservative.

[31] This is highly plausible because conservatism leads to less convergence of opinions and can thereby help "bracket" the truth. Hence, conservatism harms individual guesses, but works against the negative effect of social influence that was uncovered in Lorenz et al. (2011).

6 Conclusions

6.1 Summary and Conclusions

An organization's fit to the environment depends on the management's ability to assess the state of the – usually dynamic – environment and to cope with uncertainty. We measure team performance in this respect by assessing its ability to estimate correct answers to factual questions. Having a team leader who is knowledgeable or confident in a given topic can in principle be helpful. However, our experimental results show that communicating the leader's qualities can undermine this effect. Stressing the expertise or confidence of the leader triggers other team members to put too much weight on the leader's opinion. This narrows the opinion space and diminishes the wisdom of the group substantially. Past accuracy (T1) and actual ability are correlated such that there is a positive effect of an accurate leader, which, however, is immediately undermined by the effect of declaring it. Past confidence (T2) is only weakly correlated with actual ability such that the net effect is negative. In addition, most people suffer from overprecision, a form of overconfidence that leads to conservatism in updating and hence ignorance of the others' valuable opinions. The two effects together imply that the own and the leader's opinion are heavily weighted at the expense of the other group members' opinions, resulting in an information loss.

We investigate the opinion dynamics by looking at different classes of learning models. In particular, rational learning models in which social learning is efficient, independent of the team leader, fall short of explaining our data. A better fit is obtained for naïve learning models, which predict that the leader is more influential than any other team member. Among those, the model that gives tremendous weight to the leader (*Corazzini et al. Model*) does not fit well in the random treatment T0, but fits particularly well in the treatments T1 and T2, in which the leader is not selected at random. Compared to all models, people tend to adapt too little to the others' opinions and are too confident in their own subjective estimates.[32] To introduce this pattern in the theory of social learning, we extend both rational and naïve models by conservatism. With this twist, the fit of each model to the data increases substantially, despite the fact that the model predictions move further away from the correct answers.

Given these results, the individually optimal updating rule is a complex matter: The optimal weight on the own opinion does not only depend on the distribution of expertise in the team, but also on the behavior of the other group members. In particular, if a team leader adequately aggregates the information of the team, a team member's conservatism prevents him from learning from the others, but if a team leader inadequately aggregates the information of the team, for instance because she is confident and does not listen to the other team members, then it is very difficult for a team member to learn from the other team members. We observe

[32]The substantial amount of conservatism that we find in this paper can be partially due to the more realistic setup with the lack of common knowledge about the others' signal precisions.

that the loss in efficiency in the confidence treatment stems from both sources, team leader's behavior and followers' behavior: Confident team leaders do not sufficiently take into account the valuable opinions of others; members of such a team give her a too high weight.

One conclusion is that our paper shows advantageous effects of random leader selection ("sortition"). This political power selection rule has its roots in ancient Greece and has been discussed by various names such as "demarchy" or "aleatory democracy" (Zeitoun et al., 2014; Frey and Osterloh, 2016). While there have been discussions in the literature about the advantages and disadvantages of aleatory democracy, there is hardly empirical evidence. Our empirical results demonstrate that random selection may be beneficial compared to selection based on confidence, because then, the leader's guesses get less and other group members' guesses get more weight, taking more advantage of all information available to the team. In reality, competence is often hard to measure. When the selection criterion and competence are only weakly correlated, the leader's opinion is likely to be overrated.

6.2 Limitations

The advantages of our experimental design come at the expense of certain limitations. First, we focus on the ability of participants to learn from each other such that they find good answers to estimation questions. However, sometimes it is less important to accurately assess the environment, but to converge towards a common opinion. This may reduce conflicts and helps to work on the same tasks and to support each other. For example, it has been shown that a leader's overprecision, or resoluteness, can foster coordination and cohesion (Bolton et al., 2013). Hence, there is a trade-off between strong leadership and information loss. Second, our experimental design focuses on social learning and does not mix it with the decision-making process. Adding a decision-making stage (e.g., with a voting procedure), would increase the experiment's scope but distort measures for social learning, because participants would anticipate the decision-making stage in the social learning stage. Third, by studying star networks, we have not varied the network architecture, but only the network positions, which for star networks boils down to the question of who is the leader. Follow-up research might include a variety of network architectures and even consider endogenous network formation. Finally, the external validity of this type of experiments depends on whether the interaction among participants (who were virtually all university students) is sufficiently related to the interaction among members of real teams in organizations. We have exogenously varied the selection criterion of the leader. This resembles the perspective of the top management, deciding about, e.g., the promotion criteria of more or less senior employees of the organization.

6.3 Practical Implications

Despite these limitations of our experiment, our findings do suggest several practical implications. First, when selecting a leader, self-confidence is a dangerous proxy for competence. In

fact, real competence might be difficult to measure. What could be more easily assessed by some guessing tasks is a candidate's degree of overprecision, which might be more predictive for team learning. Second, the way the selection criterion for the leader is communicated to a team heavily affects the team's interaction and performance. In particular, stressing that the team leader was selected because of her (alleged) superiority increases her power, which might push team learning out of balance. Third, we can validate that communication and social influence can be harmful for the wisdom of crowd effect (Lorenz et al., 2011), as the crowd error increases over time. However, and importantly, we also show that social influence can foster social learning. In particular, the individual error and the collective error improve over time. Hence, interaction is not generally harmful.

Crucially, the effect of social influence on performance is moderated by the selection criterion of who is in the powerful position in the communication network, and by the declaration of the selection criterion. In conclusion, if teams want to utilize the wisdom of crowds within their team, our results suggest that they should admit interaction and opinion exchange to counter conservatism, but prevent central individuals from becoming overly influential.

References

Acemoglu, Daron, Kostas Bimpikis, and Asuman Ozdaglar. 2014. "Dynamics of information exchange in endogenous social networks." *Theoretical Economics* 9:41–97.

Acemoglu, Daron, Asuman Ozdaglar, and Ali ParandehGheibi. 2010. "Spread of (mis)information in social networks." *Games and Economic Behavior* 70:194–227.

Ambuehl, Sandro and Shengwu Li. 2018. "Belief updating and the demand for information." *Games and Economic Behavior* 109:21–39.

Aumann, Robert J. 1976. "Agreeing to disagree." *The annals of statistics* pp. 1236–1239.

Battiston, Pietro and Luca Stanca. 2015. "Boundedly rational opinion dynamics in social networks: Does indegree matter?" *Journal of Economic Behavior & Organization* 119:400–421.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch. 2014. "hroot: Hamburg registration and organization online tool." *European Economic Review* 71:117–120.

Bolton, Patrick, Markus K. Brunnermeier, and Laura Veldkamp. 2013. "Leadership, coordination, and corporate culture." *The Review of Economic Studies* 80:512–537.

Brandts, Jordi, Ayça Ebru Giritligil, and Roberto A Weber. 2015. "An experimental study of persuasion bias and social influence in networks." *European Economic Review* 80:214–229.

Çelen, Boğaçhan, Shachar Kariv, and Andrew Schotter. 2010. "An experimental test of advice and social learning." *Management Science* 56:1687–1701.

Chandrasekhar, Arun G, Horacio Larreguy, and Juan Pablo Xandri. 2015. "Testing models of social learning on networks: Evidence from a lab experiment in the field." Technical report, National Bureau of Economic Research.

Choi, Syngjoo, Douglas Gale, and Shachar Kariv. 2005. "Behavioral aspects of learning in social networks: an experimental study." *Advances in Applied Microeconomics* 13:25–61.

Corazzini, Luca, Filippo Pavesi, Beatrice Petrovich, and Luca Stanca. 2012. "Influential listeners: An experiment on persuasion bias in social networks." *European Economic Review* 56:1276–1288.

DeGroot, Morris H. 1974. "Reaching a Consensus." *Journal of the American Statistical Association* 69:118–121.

DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel. 2003. "Persuasion Bias, Social Influence, And Unidimensional Opinions." *The Quarterly Journal of Economics* 118:909–968.

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

Frey, Bruno S. and Margit Osterloh. 2016. "Aleatoric Democracy." Technical report, CESifo Group Munich.

Friedkin, Noah E. 1991. "Theoretical Foundations for Centrality Measures." *The American Journal of Sociology* 96:1478–1504.

Friedkin, Noah E. and Eugene C. Johnsen. 1990. "Social influence and opinions." *Journal of Mathematical Sociology* 15:193–206.

Gale, Douglas and Shachar Kariv. 2003. "Bayesian learning in social networks." *Games and Economic Behavior* 45:329–346.

Gervais, Simon and Itay Goldstein. 2007. "The positive effects of biased self-perceptions in firms." *Review of Finance* 11:453–496.

Golub, Benjamin and Matthew O. Jackson. 2010. "Naïve Learning in Social Networks and the Wisdom of Crowds." *American Economic Journal: Microeconomics* 2:112–49.

Golub, Benjamin and Matthew O. Jackson. 2012. "How homophily affects the speed of learning and best-response dynamics." *The Quarterly Journal of Economics* 127:1287–1338.

Grimm, Veronika and Friederike Mengel. 2018. "An Experiment on Belief Formation in Networks." *Journal of the European Economic Association* .

Haslam, S. Alexander, Craig McGarty, Patricia M. Brown, Rachael A. Eggins, Brenda E. Morrison, and Katherine J. Reynolds. 1998. "Inspecting the emperor's clothes: Evidence that random selection of leaders can enhance group performance." *Group Dynamics: Theory, Research, and Practice* 2:168–184.

Herz, Holger, Daniel Schunk, and Christian Zehnder. 2014. "How do judgmental overconfidence and overoptimism shape innovative activity?" *Games and Economic Behavior* 83:1–23.

Keuschnigg, Marc and Christian Ganser. 2017. "Crowd Wisdom Relies on Agents' Ability in Small Groups with a Voting Aggregation Rule." *Management Science* 63:818–828.

Lorenz, Jan, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. "How social influence can undermine the wisdom of crowd effect." *Proceedings of the National Academy of Sciences* 108:9020–9025.

Mannes, Albert E. 2009. "Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision." *Management Science* 55:1267–1279.

Mannes, Albert E. and Don A. Moore. 2013. "A Behavioral Demonstration of Overconfidence in Judgment." *Psychological Science* 24:1190–1197.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." Technical report, National Bureau of Economic Research.

Moore, Don A. and Paul J. Healy. 2008. "The trouble with overconfidence." *Psychological Review* 115:502–517.

Moussaïd, Mehdi, Juliane E. Kämmer, Pantelis P. Analytis, and Hansjörg Neth. 2013. "Social Influence and the Collective Dynamics of Opinion Formation." *PLOS ONE* 8:1–8.

Mueller-Frank, Manuel. 2013. "A general framework for rational learning in social networks." *Theoretical Economics* 8:1–40.

Peterson, Cameron R. and Lee R. Beach. 1967. "Man as an intuitive statistician." *Psychological Bulletin* 68:29.

Phan, Tuan, Adam Szeidl, and Markus Mobius. 2015. "Treasure Hunt: A Field Experiment on Social Learning." mimeo, Society for Economic Dynamics.

Rauhut, Heiko and Jan Lorenz. 2011. "The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions." *Journal of Mathematical Psychology* 55:191–197.

Rosenberg, Dinah, Eilon Solan, and Nicolas Vieille. 2009. "Informational externalities and emergence of consensus." *Games and Economic Behavior* 66:979–994.

Soll, Jack B. and Joshua Klayman. 2004. "Overconfidence in interval estimates." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30:299.

Surowiecki, J. 2004. *The Wisdom of Crowds*. New York: Random House.

Zeitoun, Hossam, Margit Osterloh, and Bruno S. Frey. 2014. "Learning from ancient Athens: Demarchy and corporate governance." *The Academy of Management Perspectives* 28:1–14.

A Appendix: Additional Tables and Figures

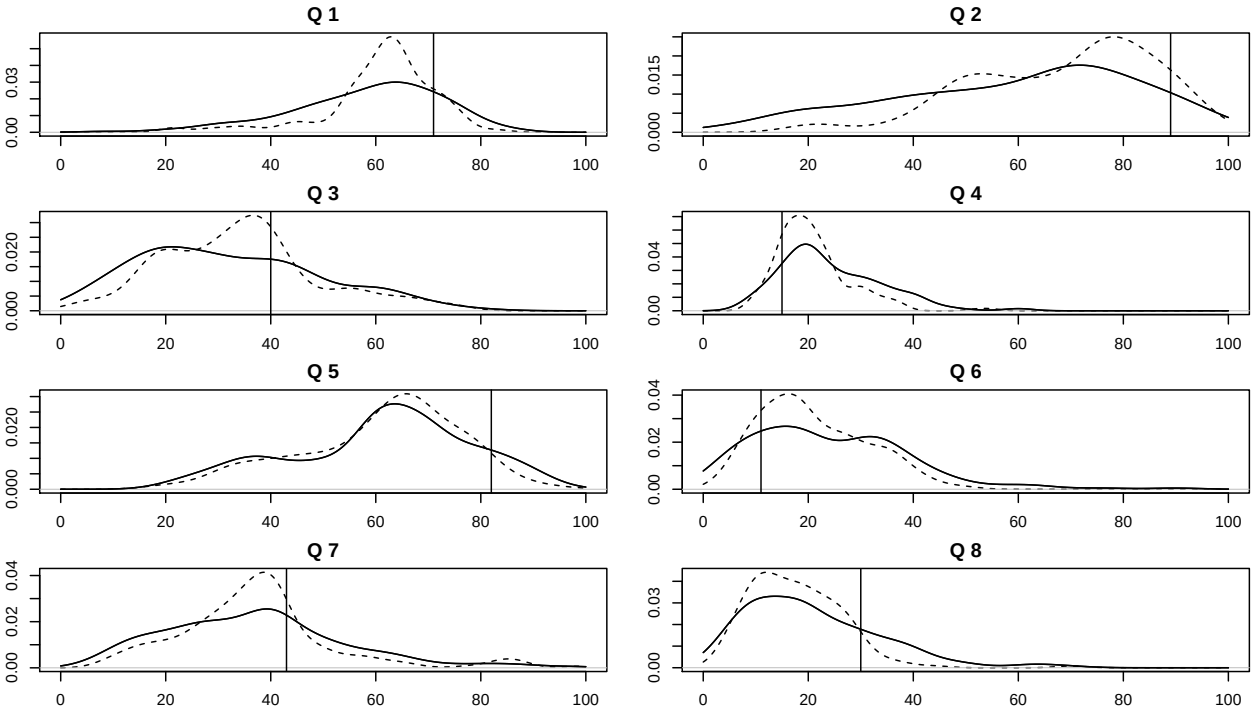

Figure A.1: Distribution of answers by questions. Black lines show the first round ($t = 1$); dashed lines show the last round ($t = 6$). The correct answer is indicated by the vertical line.

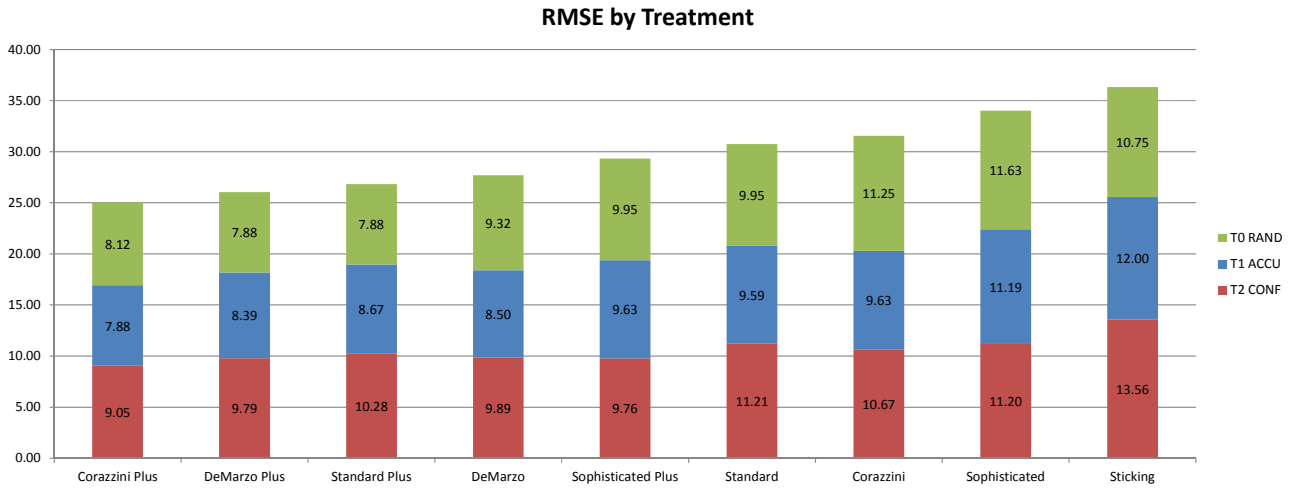

Figure A.2: Root mean squared errors (RMSE) of social learning models differentiated by treatment. Lower errors mean better fit between model and data.

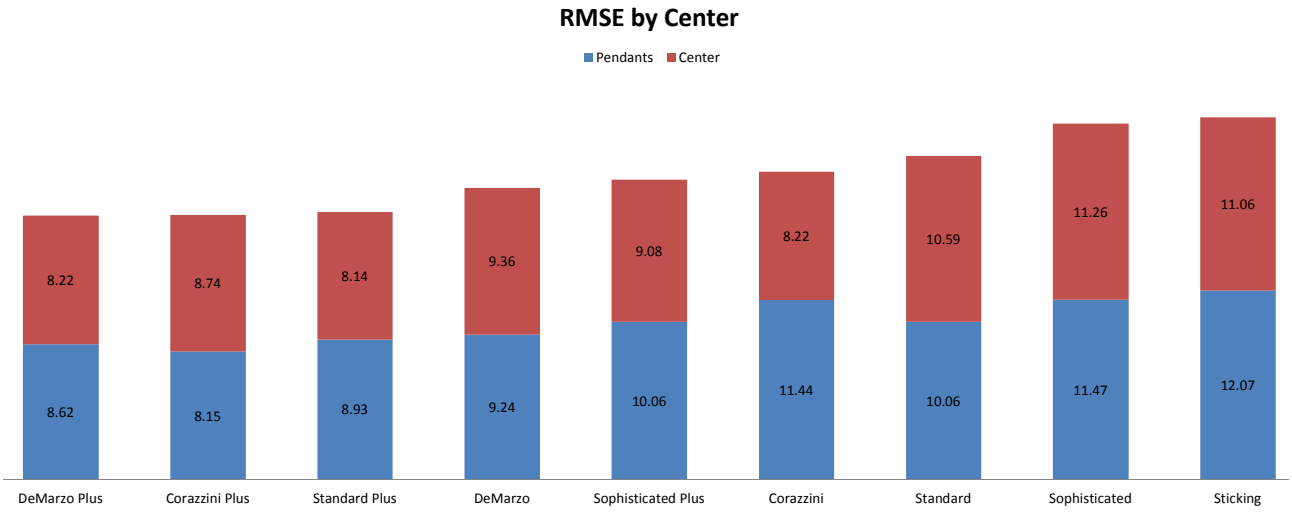

Figure A.3: Root mean squared errors (RMSE) of different models by center and pendants differentiated by center and pendants. Lower errors mean better fit between model and data.

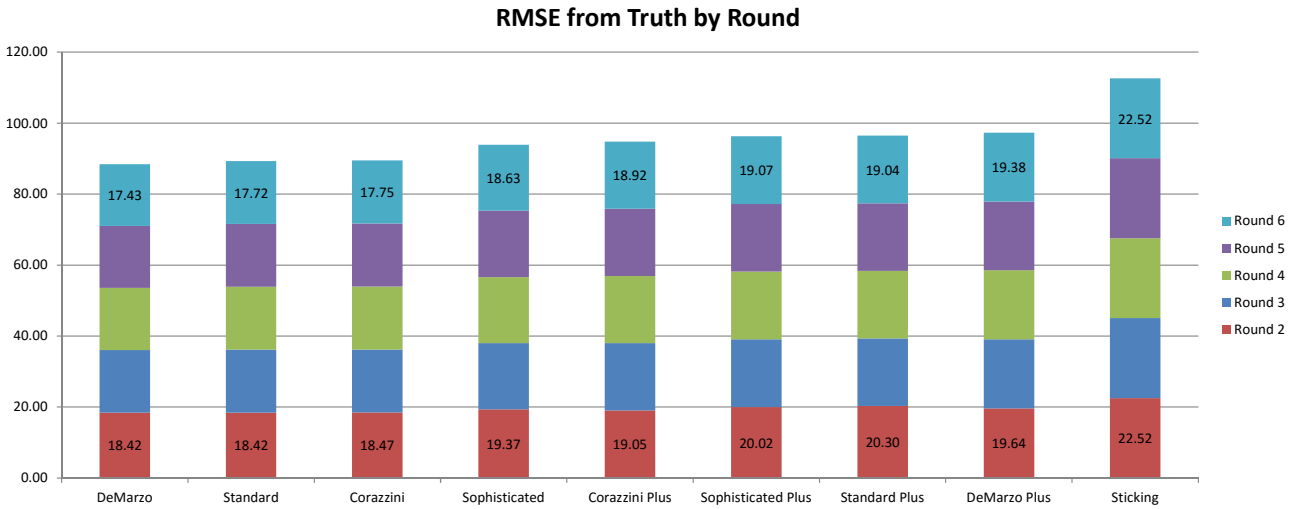

Figure A.4: Root mean squared errors (RMSE) of social learning models to the correct answer. "Standard" and "Sophisticated" are models of rational learning; "DeMarzo" and "Corazzini" are models of naïve learning. "Plus" models incorporate conservatism. Lower errors mean better fit between model and truth.

Note: Restricting attention to the random treatment T0, the order of models according to their fit to the truth is as follows: *Standard Model* 18.01, *DeMarzo et al. Model* 18.22, *Sophisticated Model* 18.83, *Corazzini et al. Model* 18.92, *Sophisticated-Plus Model* 19.28, *Standard-Plus Model* 19.35, *Corazzini et al. Plus Model* 19.59, *DeMarzo et al. Plus Model* 19.80, *Sticking Model* 22.51.

	(1) Exp. Payoff	(2) Exp. Payoff	(3) Exp. Payoff
accuracy-trait		0.125	0.102
		(1.23)	(1.04)
accuracy-declaration (T1)	-0.0900	-0.193	-0.189
	(-1.02)	(-1.57)	(-1.59)
confidence-trait		0.126	0.186
		(1.33)	(1.93)
confidence-declaration (T2)	-0.173*	-0.261*	-0.286*
	(-2.05)	(-2.30)	(-2.56)
overprecision center			-0.0704**
			(-3.10)
overprecision pendants (avg.)			-0.0909*
			(-2.37)
intercept	0.478***	0.412***	1.092***
	(5.85)	(5.05)	(5.44)
N	1'408	1'408	1'408

Question dummy coefficients for 8 questions not shown.

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.1: Treatment effects on expected payoff in EUR (for period 6). Linear regressions.

	(1) T0 random	(2) T1 accuracy	(3) T2 confidence
own weight (pendant)	0.567***	0.405***	0.392***
	(16.23)	(9.90)	(9.40)
center's weight	0.267***	0.449***	0.469***
	(7.86)	(11.64)	(8.10)
other pendants' weight	0.166***	0.146***	0.139**
	(5.37)	(3.92)	(3.28)
N	528	264	264

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.2: Influence weights on pendants' final answer, separately estimated for each treatment. Regression of the pendant's final answer (period 6) on the initial answers (period 1). Coefficients forced to sum up to one.

	(1) T0 random	(2) T1 accuracy	(3) T2 confidence
own weight (center)	0.473***	0.659***	0.705***
	(9.04)	(10.54)	(9.23)
pendants' weight	0.527***	0.341***	0.295***
	(10.06)	(5.46)	(3.86)
N	176	88	88

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.3: Influence weights on center's final answer, separately estimated for each treatment. Regression of the center's final answer (period 6) on the initial answers (period 1). Coefficients forced to sum up to one.

	(1) pendant's answer_6 (last period)
own weight (pendant)	0.577***
	(13.77)
center weight	0.244***
	(5.51)
other pendants' weight	0.198***
	(4.78)
accuracy-trait × own	-0.0234
	(-0.41)
accuracy-trait × center	0.0693
	(1.30)
accuracy-trait × other pendants	-0.0393
	(-0.82)
accuracy-declaration (T1) × own	-0.140*
	(-2.04)
accuracy-declaration (T1) × center	0.120*
	(2.02)
accuracy-declaration (T1) × other pendants	0.0222
	(0.38)
confidence-trait × own	-0.00712
	(-0.12)
confidence-trait × center	0.0317
	(0.68)
confidence-trait × other pendants	-0.0516
	(-1.04)
confidence-declaration (T2) × own	-0.152*
	(-2.23)
confidence-declaration (T2) × center	0.169**
	(2.64)
confidence-declaration (T2) × other pendants	0.0407
	(0.70)
N	1'056

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.4: Influence weights on pendant's final answer. Linear regression of the pendant's final answer (period 6) on the initial answers (period 1).

	(1) center's answer_6 (last period)
own weight (center)	0.400***
	(6.30)
pendants weight	0.643***
	(9.95)
accuracy-trait × own	0.158*
	(2.17)
accuracy-trait pendants	-0.147
	(-1.93)
accuracy-declaration (T1) × own	0.0402
	(0.44)
accuracy-declaration (T1) × pendants	-0.0393
	(-0.38)
confidence-trait × own	0.139*
	(1.97)
confidence-trait × pendants	-0.189*
	(-2.52)
confidence-declaration (T2) × own	0.108
	(1.28)
confidence-declaration (T2) × pendants	-0.0353
	(-0.38)
N	352

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.5: Influence weights on center's final answer. Linear regression of the center's final answer (period 6) on the initial answers (period 1).

	(1) individual error (log)	(2) collective error (log)	(3) crowd error
accuracy-trait	-0.0989	-0.0589	0.0143
	(-1.02)	(-0.56)	(0.04)
accuracy-declaration (T1)	0.108	0.0709	0.185
	(0.95)	(0.58)	(0.50)
confidence-trait	-0.136	-0.264*	-0.695*
	(-1.48)	(-2.43)	(-2.35)
confidence-declaration (T2)	0.238*	0.355**	1.215**
	(2.17)	(2.83)	(3.22)
overprecision center	0.0426	0.0453	0.216**
	(1.93)	(1.92)	(3.17)
overprecision pendants (avg.)	0.0804	0.0803	0.305*
	(1.94)	(1.73)	(2.09)
intercept	1.696***	1.706***	
	(7.87)	(6.98)	
intercept cut 1			-0.637
			(-0.85)
intercept cut 2			1.154
			(1.54)
N	1'408	352	352

Question dummy coefficients for 8 questions not shown

t statistics in parentheses; robust standard errors used; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.6: Treatment effects on final errors: log individual error, log collective error and wisdom of crowd error (in period 6). Linear regression (models 1 and 2) and ordered logit regression (model 3). For models 1 and 2 the null cannot be rejected for the overprecision coefficients (at the 5% significance level). The corresponding p values are 0.054 (overprecision center) and 0.053 (overprecision pendants) in model 1; and 0.056 (overprecision center) and 0.084 (overprecision pendants) in model 2.