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, Economics, berno.buechel@unifr.ch

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

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

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

June 5, 2018

Abstract

We investigate how the selection process of a leader affects team performance with respect to social learning. We use a lab 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. If anything, they even outperform teams with leaders selected on self-confidence. Hence, self-confidence can be a dangerous proxy for competence of a leader. We can show that it is the declaration of the selection procedure which makes non-random leaders overly influential. We set up a horse race between several rational and naïve models of social learning to investigate the micro-level mechanisms. We find that overconfidence and conservatism contribute to the fact that overly influential leaders may mislead their team in finding good estimates.

JEL-Code: D83, D85, C91.

Keywords: Social Networks, Social Influence, Confidence, Overconfidence, Bayesian Updating, Naïve Learning, Sortition, 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.

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 upon using the expertise of a single individual by harnessing the wisdom of crowds (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 the initial level of information of each team member, the accuracy of the updated opinions 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* in a lab experiment. 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 questions, we set up a lab experiment in which subjects are asked to answer incentivized estimation questions repeatedly. After each round, subjects can observe the guesses and the confidence levels of some of their team members according to a star network with the leader at the center. Thus, every team member observes the guesses of the leader, while only the leader observes the guesses of all members. We randomly allocate subjects into three treatments, which differ by the criterion that determines how the team leader is selected. In the baseline treatment (T0), the leader, i.e., the center, is selected at random. In the accuracy treatment (T1), the leader is the group member whose estimation of a related question was the most accurate in the team. Finally, in the confidence treatment (T2), the team member with the highest stated level of confidence (in the own answer of a related question) is selected. Potential ties in maximal accuracy or maximal confidence are broken at random.

Interestingly, a set of theoretical models following from the Bayesian approach to social learning predict for this setting that the selection of the center does not matter for the outcome, apart from the first two rounds, and that social learning is efficient.[2] The reason is that agents can exchange their opinions such that proper aggregation leads to a common estimate (consensus) that is independent of who is at the center of the communication network. In

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

contrast, a set of models of naïve social learning predict a strong impact of the center on the same process, which induces an inefficient outcome.[3] Based on the assumption that subjects fail to account correctly for the repetition of the center's and the others' initial opinion, they predict that consensus is approached over time, but with a strong "bias" towards the center's initial opinion. In particular, the center's weight on the consensus opinion is predicted to be proportional to her eigenvector centrality, which is several times larger than the other team members', in standard specifications of a naïve model of social learning. Unless the leader is much better infomed than the other team members, this is suboptimal, giving the leader's opinion too much weight. Now, 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 declaration of why the leader was selected.

Results. In the experiment, 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. Our first result is that leader selection based on accuracy (T1) or confidence (T2) does not outperform the random selection. Leader selection based on confidence (T2) may even undermine performance. The reason for this insight becomes apparent when isolating the effect of declaring how the leader is selected. The declaration of the leader as somewhat superior, be it in terms of past performance (T1) or of confidence (T2), induces the other team members to put more weight on the leader's opinion, making the team vulnerable to be misled by a single person. In contrast, teams with random leaders more equally weight each other's opinions with the consequence of a higher performance in our setting. On top of these effects, we assess how team performance is affected by (judgmental) 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). It turns out that both overconfident leaders and overconfident other team members undermine performance. Hence, when designing a procedure for leader selection in a situation in which social learning is important, declared random selection is a viable option and overconfidence should be avoided.

In the second part of the paper, we set up a horse race between different models of social learning to shed more light on individual learning behavior and on the mechanisms of how leader selection affects the wisdom of crowds in networks. Despite a long tradition of theoretical insights and a growing body of empirical research, social learning behavior is still far from being fully understood. In line with the previous literature, we observe that simple models (of naïve social learning) generally fit better than sophisticated models (of Bayesian social learning). This has the consequence that the leader's weight on the long-term opinion is large already due to her central position in the network structure. Moreover, the experimental data reveal that an important pattern is missing in both theoretical approaches: People tend to adapt their opinion

[3] 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.

less than predicted, a pattern called *conservatism*. Conservatism is a very common finding in experiments on belief updating and can be caused by (judgmental) overconfidence, as we argue in this paper.[4] We incorporate this feature, which is largely missing in the theoretical literature on social learning, into both model classes and observe that incorporating conservatism improves the model fit of both model classes. Hence, there is important feedback from our data to theory development. Incorporating conservatism is a behavioral twist that matches empirical findings and also affects performance of social learning.

Methodological Approach. In laboratory experiments (and in lab in the field experiments), theoretical models can be directly tested. For instance, Corazzini et al. (2012), Grimm and Mengel (2016), and Chandrasekhar et al. (2015) compare sophisticated models of Bayesian learning with simple models of naïve learning in settings in which their predictions diverge. The common conclusion is that the observations are more often consistent with the simple models. Similarly, Choi et al. (2005) and Çelen et al. (2010) study predictions of social learning models experimentally. A caveat of this theory-testing approach is that the participants are confronted with highly stylized tasks such as guessing an average (or its sign) of randomly drawn numbers (Corazzini et al., 2012; Çelen et al., 2010) or finding an abstract true state (Choi et al., 2005; Grimm and Mengel, 2016; Chandrasekhar et al., 2015). It is questionable how the investigated learning behavior transfers to settings with real questions. A lab in the field approach (Chandrasekhar et al., 2015) does not fully mitigate this issue of external validity, because the types of questions are still often stylized. At the other side of the spectrum, real teams could be studied in the field to address our main research question (without the issue of external validity). However, besides the issues of noise, missing values, and the problem to measure performance of social learning, there would be a severe endogeneity problem. First, because face-to-face interaction gives rise to effects (e.g., due to charisma), which are difficult to control for, and second, because there is usually no proper randomization on who becomes a leader. For these reasons, we decided to use some middle ground between these two approaches (theory-testing experiment and field data) in order to complement them. For that purpose, we imported a method developed outside of economics which has been increasingly used recently (Lorenz et al., 2011; Rauhut and Lorenz, 2011). 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). The questions cover various topics and create a natural uncertainty among the participants who are paid according to their answers' accuracy. Arguably, teams who are able to estimate such factual questions accurately are also better at estimating states that cannot be simply measured, or at estimating future states of the world

[4] Experiments on belief updating frequently find that real people are more conservative updaters than the theoretical model would predict (Mobius et al., 2011; Ambuehl and Li, 2018; Mannes and Moore, 2013), 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[.]"

(which is of high relevance in real managerial or political teams). In our experiment, however, the quality of social learning can be assessed without waiting for the future to realize. The slightly added realism of that approach already changes the way subjects communicate with each other because, given that there is no stylized draw of signals which is common knowledge, it becomes important not only to communicate the guess, but also the own confidence in the guess. We consider it as a realistic assumption that people can "tag" the pieces of information they pass on with a confidence level by stating how confident they feel about their own guess.[5] This 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.

Contribution. Our paper entails three contributions. First, we provide some empirical evidence for the advantage of a selection procedure that is based on random leader selection ("sortition"). For both corporate and political governance, *sortition* (also called demarchy, allotment, or aleatory democracy) is discussed as an alternative selection procedure, which has its roots in ancient Athens and medieval Italy (Zeitoun et al., 2014; Frey and Osterloh, 2016). Despite a long list of claimed advantages of this procedure, empirical evidence showing its superiority is very rare. 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.[6] Our results show that random selection can also be beneficial in other settings. We additionally demonstrate that it is the declaration of randomness rather than the selection at random per se that is the crucial aspect here. The strength of random selection is based on the fact that the leader's influence on team members is not amplified by declaring the leader's specialty. Since the leader is already special because of her network position, additionally highlighting the leader's properties by declaring them to be relevant for selecting the leader makes the others prone to insufficiently weigh their own opinions, which may result in a loss, because the wisdom of crowds in the group is not harnessed.

Second, our experimental data reveal that the extent of (judgmental) overconfidence, i.e., providing too narrow confidence intervals, has a deteriorating effect on team performance. Indeed, both overconfident leaders and overconfident team members undermine performance. For the selection of leaders within organizations, this suggests that either overconfident leaders should be generally avoided or that there is at least a trade-off between beneficial effects of a leader's overconfidence (e.g., to foster coordination, Bolton et al. 2013, or to motivate team members, Gervais and Goldstein 2007) and the negative effect on social learning. (Judgmental)

[5] This is similar to the literature that considers "tagging" pieces of information with their source (Acemoglu et al., 2014; Phan et al., 2015).

[6] 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.

overconfidence may be partially domain-specific and state-dependent, but to some extent it is a personality trait that can easily be assessed, e.g., in an assessment center in the course of a selection procedure.

Third and finally, our paper makes a methodological contribution. By combining the experiments on factual questions with the theories on social learning, we bridge between neat theoretical frameworks and experimental set-ups that are less stylized (than those used for pure theory testing). By building this bridge it becomes apparent that the assumption of common knowledge about signal precision is problematic. Arguably, in reality people do not know the signal precision of their interaction partners, but form expectations about it, given what they know about this person and given how this person "tagged" her piece of information with a level of confidence. (Judgmental) overconfidence, as well as mistrust or anchoring effects, can lead to *conservatism* in updating, i.e., agents incorporate new pieces of information less than theoretically predicted. Bolton et al. (2013) further argue that other behavioral biases such as a selection bias in information acquisition can also induce conservatism (what they call resoluteness). We incorporate this idea into both naïve and rational models of social learning and find that the model fit of each model increases. This is informative for economic theory on naïve and rational social learning by opening an avenue for an empirically important model extension. In particular, our simple extensions of the models alter the prediction that consensus is reached or approached. Instead, they predict a persisting diversity of opinions, in which each agent's long-run opinion is "biased" in the direction of his initial opinion. This qualitatively different prediction could be studied more generally and be tested in follow-up studies.

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] Participants were mostly undergraduate students from various disciplines; there was no restriction on the pool of participants. In each session, participants were randomly allocated into groups of four. The basic task was to answer a factual question individually and to provide 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.4. On average, sessions lasted for one hour and participants earned 9.50 Euros, which was close to the 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

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

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 lead to expulsion from the experiment and exclusion from all payments.

Each session consisted of two phases: A selection phase I and a decision phase II. In the selection phase I, each participant answered eight different factual questions once. At the end of the experiment, one of these questions was randomly selected to be payoff-relevant. In the decision phase II, there was another set of eight questions, each of which was similar to one of the questions of the selection phase. 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 related to diverse topics and each question was already tested in previous experiments (Lorenz et al., 2011; Rauhut and Lorenz, 2011; Moussaïd et al., 2013).[8]

In the decision phase II, each question had to be answered six times in a row, in a sequence of six rounds. After each round, participants received feedback about the answers and confidence statements provided by their group members according to a star network. The center of the star network could observe the previous answers and confidence levels 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 center became the group member whose guess on the similar question in phase I was closest to the correct answer. In the confidence treatment T2, the center became 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. Half of all groups played the random treatment (T0) for four questions and the accuracy treatment (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. When the network for one question was formed, the selection procedure was made transparent to the group members. In the selection phase I, subjects did not know how decisions in the selection phase could have an influence on the decision phase. 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 center or to avoid becoming the center 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. As we discuss in the next section, among rational agents there are indeed incentives to communicate truthfully the level of confidence in

[8] The full list of questions can be found in Online Appendix C.

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.

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. There is an unkown state of the world $\theta \in \Theta$, which is the correct answer to the question at hand.[9] 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 $x_1(2)$. 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.[10]

[9] 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 $\Theta = [0, 1]$.

[10] A formal statement of this result can be found in Online Appendix B. There we introduce the general framework (B.1.1), prove the proposition (B.1.2), and provide two specific examples how such a rational model

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 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 states the payoff-maximizing guess, which implies that social learning is "efficient" in the sense of maximizing the sum of expected payoffs. However, several of its underlying assumptions deserve further attention.

First, a rational agent i is assumed to state the answer $x_i(t)$ that maximizes expected payoff, given his belief. This holds at least in the last round $t = 6$. In earlier rounds, there is potentially a strategic incentive to provide an answer that does not maximize expected payoff of that round (in order to be able to provide a better answer in later rounds). In fact, the earliest possibility to realize a deviating strategy is to deviate in round $t = 1$, learn something about the reaction of others in round $t = 2$, and materialize the better guess in round $t \geq 3$. Since, under the assumptions above, each agent states the optimal answer from round $t = 3$ on, strategic misrepresentation cannot pay off. There is simply no room for improvement. The same argument applies to the strategic misrepresentation of confidence statements. Hence, strategic misrepresentation is not an issue in our setting.

Second, 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.[11] 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.

Third, 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

unfolds in our setting (B.2.1).

[11] We study such models in Section B.2.1.

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.[12]

Fourth 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, the Prediction 1 gives us a clean baseline to compare the data with.

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; Grimm and Mengel, 2016; Battiston and Stanca, 2015; Chandrasekhar et al., 2015). 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 $x_i(6)$. It can be shown that, for any i,

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

[12] 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.

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.[13]

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 structure, 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, the 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 5.3.

Our treatments T1 and T2 mainly affect naïve social learning through the manipulation of the network structure (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

[13] A formal statement of this result can be found in Online Appendix B.1.3. There we introduce a probabilistic framework and prove the proposition.

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.

4.1 Performance over Time

We measure the quality of the final answers both on the individual and on the collective level. On the individual level, we measure the quality by the error $e_i(t)$, which is the absolute distance between answer $x_i(t)$ and truth θ. On the group level, we use two complementary measures. First, we measure the quality of the four answers by the *collective error*, i.e., the error of the mean of the four answers in the group $ce(t) = |\frac{1}{4}\sum_{i=1}^{4} x_i(t) - \theta|$. Indeed, given the four final answers by a group, a decision might be taken on the basis of the mean of the four answers. Second, we consider 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. We define the indicator variable *(wisdom of) crowd error* as follows: $woce(t) = 0$ 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. The crowd error measures the error made when assuming that the correct answer lies between the given answers. For all three measures of performance, smaller errors mean higher performance.

Figure 1 depicts the levels of these performance measures over time, distinguishing 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 tend to decrease over time. More precisely, the pendants' average error reduces three times significantly on the five percent level. 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). The centers' average error reduces significantly once in the random treatment T0, as well as in the confidence treatment T2, but never so in the accuracy treatment T1 (at significance levels $p < 0.05$). By and large, these observations on the learning dynamics are consistent with the predictions of the rational model, namely that pendants learn twice and centers learn once. In particular, in the random treatment T0 the center reduces her error drastically from the first round to the second without significant further improvements, as the rational model would predict. Panels D-F show that the collective errors are on average between 12 and 16 percentage points from the true answer and also reduce over time. Similarly to the individual errors, the collective errors first decrease and then seem to settle after a few rounds (at a point that is significantly greater than zero). Taking these observations on individual and collective errors together, agents do learn from each other, but most of learning takes place in the first and in the second round of updating, i.e., until round $t = 3$.[14] A similar pattern, albeit with a 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 $t = 3$. This observation is consistent with findings of Lorenz et al. (2011), who show that the exchange of opinions reduces the wisdom of crowds. Crowd error is an indicator variable of which the averages have to be interpreted correspondingly. For instance, in the random treatment T0, $woce(6)$ is 1.57 on average, which indicates that there are many cases (here: 65.9%) with a crowd error of two, and very few cases (8.5%) with a crowd error of zero. Hence, in the final period the correct answer most frequently lies outside of the convex hull of the provided answers.

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.*

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.

Table 1 reports these models when controlling for each treatment T1 and T2 with a dummy variable, while T0 is the reference category. We control for the heterogeneity between different questions by using 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

[14] Learning cannot stem from having more time to think about a question since participants of the experiment 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 participants of the experimental sessions who were not exposed to any information. We randomly selected these subjects from all participants of sessions whose number of participants was not divisible by four, the size of our groups.

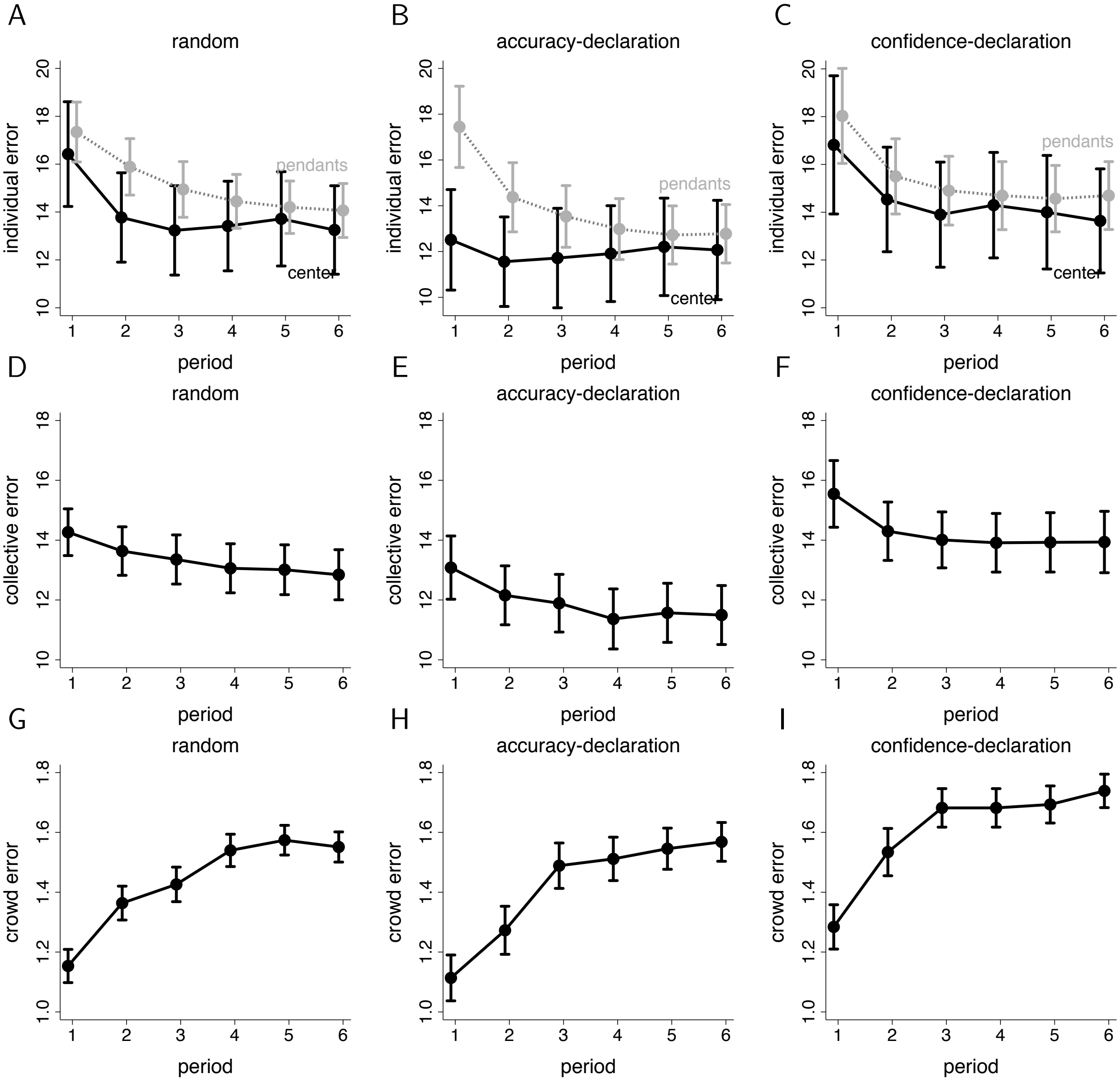

Figure 1: Individual, collective, and crowd errors over time by treatments. Panels A, B, C differentiate between centers (black) and pendants (gray). All confidence intervals are standard 95% confidence intervals.

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 1 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 the individual error ($p = 0.114$).[15] 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 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 Table A.1, model 1).

	(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 1: Treatment effects on final errors: log error, log collective error, and wisdom of crowd error (in period 6). Linear regression (models 1 and 2) and ordered logit regression (model 3).

Result 2. *Performance does not improve when the center is known to be the most accurate (T1) or the most confident (T2). If anything, performance even deteriorates when the center is known to be the most confident (T2).*

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

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. 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 or his 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 2 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, the outcome measures tend to improve, which can be seen from 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 for 0.26 EUR for this question. This is a decrease by 49% from 0.54 in the case of having the most confident in the center in the random treatment. 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.

While Table 2 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).

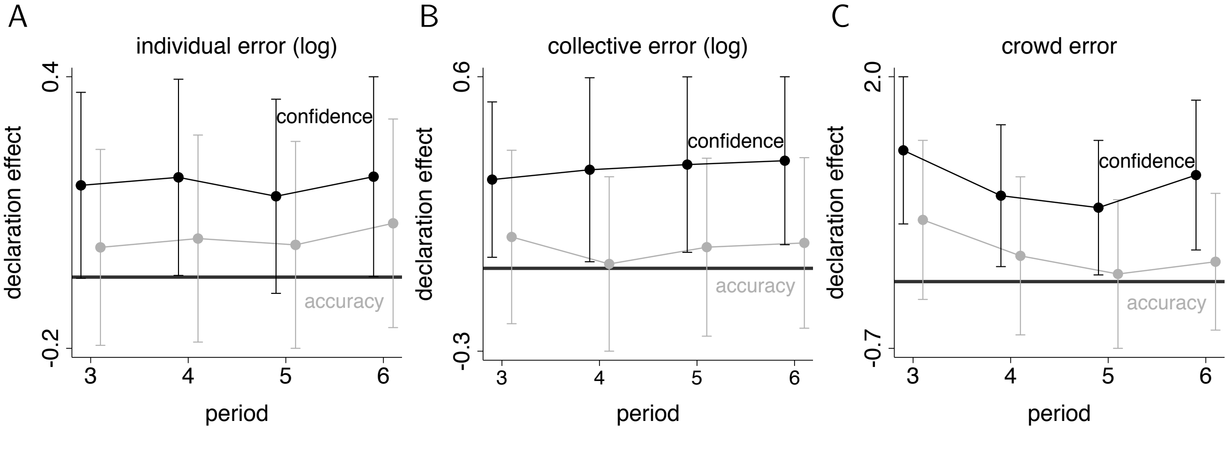

Figure 2: Treatment effects on errors: log error, log collective error, and wisdom of crowd error (periods 3-6). Linear regressions, 95 % confidence intervals.

	(1)	(2)	(3)
	individual error (log)	collective error (log)	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 2: Treatment effects on final errors: log error, log collective error, and wisdom of crowd error (in period 6). Linear regression (models 1 and 2) and ordered logit (model 3).

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.*

4.3 Social Influence

To analyze why the selection of the center can 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 in 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 when estimating them separately for each treatment. For instance, in the random treatment T0, a pendant's final answer is estimated as the linear 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 to make in these tables. First, every agent places much weight to his own initial opinion. In the rational model and the random treatment, we would expect that on average this weight is 25%.[16] Second, the weight individuals place on their own initial opinion depends on the treatment. In the random treatment, pendants place more weight on themselves than in the other two, while centers place less weight on themselves in the random treatment. 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 the center, 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. A positive effect of a certain dummy variable means that the given influence weight is increased by the given treatment.

When the center happens to be the most accurate or the most confident, but there is no public declaration of this, then the pendants do not strongly respond (panel A); if anything, they only mildly increase their weight on the center. In the same case, i.e., when the center is the most accurate or confident, 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

[16] We will return to this observation when extending the social learning models in section 5.1.

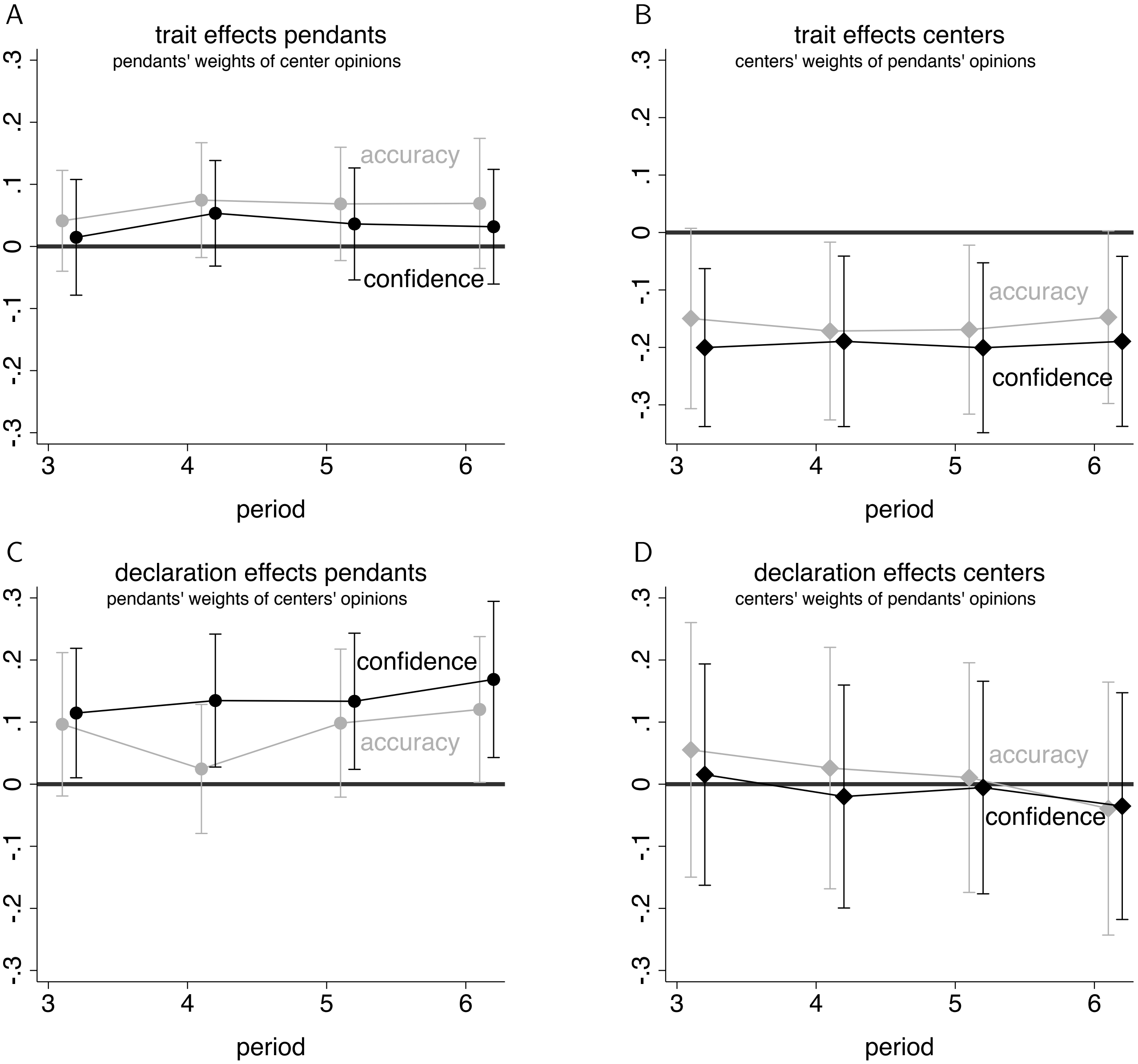

Figure 3: Trait and declaration influence for pendants and centers. Gray accuracy, black confidence treatments, 95 % confidence intervals. Panels A and C show how a pendant's answer in late periods is influenced by the center's initial answer. Panels B and D show how a center's answer in late periods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

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 the 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. Declaring that the center is somewhat special increases the weight she receives. 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, giving him/her more weight only has the negative effect of insufficiently taking into account the information of the others. This can be even worse than the random treatment T0.

4.4 Overconfidence

As we have seen in Table 2 above, it is rather beneficial for the group when the center happens to be most accurate or most confident, but is not declared as such. On the other hand, it is well-known that many people are often overconfident, 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 her individual overconfidence 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 overconfidence score in $\{0, 1, \ldots, 8\}$ with the interpretation that a person is the more overconfident the larger her overconfidence score becomes. As Figure 4 reveals, many agents are overconfident. 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 overconfidence.

In Tables A.1 and A.6 in the Appendix, we analyze how the center's overconfidence score as well as the average of the pendants' overconfidence scores impact the group's performance (on top of the previously found treatment effects). We first find that the formerly discussed effects

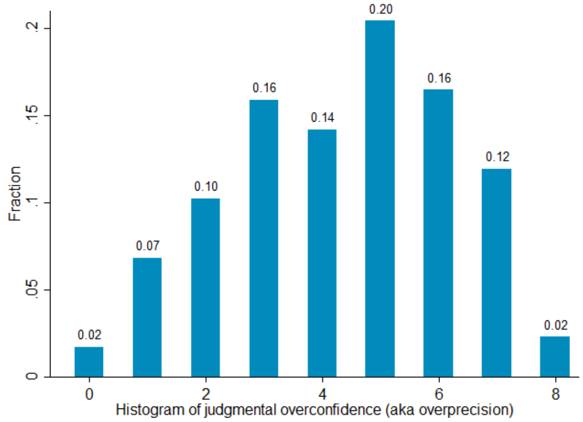

Figure 4: Histogram of individual overconfidence. The value 0 means that a subject has specified for all eight knowledge questions a respective 90% confidence interval which encloses the true value. The value 8 means that a subject has specified for all eight knowledge questions a 90% confidence interval which does not enclose 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 overconfident).

(in particular, the negative declaration effect of T2) remain significant and are hence robust when controlling for overconfidence. Second, we observe that the center's and the pendants' overconfidence coefficients are positive. 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' overconfidence undermine performance.*

Overconfidence is of course related to confidence itself, and the most confident group member 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 overconfidence 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 who should optimally be selected may well be characterized as confident, but not as being overconfident.

Hence, all results (Results 1-5) contribute to a coherent picture of how the selection of the leader affects social learning. To investigate this interpretation further, in particular the one of placing "too much weight on the center" in the confidence treatment T2, we analyze more in-depth the underlying micro-level mechanisms. In particular, we will study the fact that weights on own opinions are too large, which also prevents optimal 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.

5 Learning Behavior

The experimental data allow us to test theories of social learning on multiple levels. First, their implications for the performance of social learning (as summarized in Prediction 1 and Prediction 2) are found to be consistent with some empirical results and inconsistent with others. Second, we can directly take the theoretical models to the data and study which aspects are in line with real behavior. For this purpose, we specify and vary the models and measure which model specification best fits the data. We thus include model variations that incorporate conservatism, a pattern that is commonly found in experimental set-ups, but absent in any Bayesian model of social learning (that we are aware of) and absent in almost all naïve models of social learning.

5.1 Specification and Extension of Bayesian Models

To specify the rational models, we assume that each agent's belief follows a beta distribution. This is a standard functional form for beliefs that live on intervals.[17] With some assumptions on the distribution of signals, all agents' beliefs at any time indeed belong to the class of beta distributions.[18] Assuming conditional independence of initial signals, Bayesian agents will state guesses that are convex combinations of their initial guesses. The weight on these guesses, however, depends on the signal quality of each agent i, which we denote by n_i. The model variations that we study differ in the assumptions about signal quality.

A baseline assumption is to suppose that the precision of each agent's signal is the same, i.e., $n_i = n_j$ for all i, j. In that case, the optimal guess x^*, which will be the consensus from round $t = 3$ on, is simply the unweighted mean of the initial guesses $x_i(1)$. We call this the *Standard Model*. Alternatively, agents are assumed to communicate their belief fully by providing the guess and the confidence level. Then, for each answer $x_i(1)$ and its confidence $c_i(1)$, the center can determine the two parameters of the corresponding beta distribution and combine all initial beliefs in a rational manner, thereby updating leads to a combination of own and others' guesses – not with equal weights, but with larger weights for those guesses which are tagged by high confidence. We call this the *Sophisticated Model*. Note that these are two opposing views on the informativeness of the confidence statement – either confidence

[17] Like the normal distribution, which is a standard functional form for beliefs on the unbounded real numbers, it is determined by two parameters only.

[18] The formal framework is provided in section B.2 of the Online Appendix.

is fully informative or confidence can be ignored – which lead to two models that both satisfy the requirements of Prediction 1, and are hence similar in most respects. They differ in their weighting of initial information.

The previous empirical literature on real people's beliefs and their updating finds two very strong and consistent patterns: overprecision and conservatism.[19] There is a simple way to introduce both of them into our model: Agents overestimate their own signal precision by a factor $\tau_i \geq 1$; respectively, they underestimate the signal precision of the others by the inverse factor $\frac{1}{\tau_i}$. The motivation of this model variant is that overconfident agents suffer from overprecision in the sense that they perceive their signal as more precise than it is.[20]

Formally, this is a generalization of the *Standard Model* and the *Sophisticated Model*. This model also predicts that there are no more changes after $t = 3$. However, this model does not predict consensus! The agents' opinions settle down in between the prediction of x^* (i.e., the case $\tau_i = 1$ for all $i \in N$) and their initial guess $x_i(1)$. The weight of the own initial guess is thereby increasing in overprecision τ_i. In particular, if $\tau_i \to \infty$, then $x_i(t) \to x_i(1)$, i.e., infinitely overprecise agents are totally conservative and always stick to their initial guess. (We will include such a model as a baseline and call it the *Sticking Model*.)

To specify concrete models, we choose levels of overprecision τ_i that match with empirical results on overprecision. When asked for a 90% confidence interval, many people provide a 50% confidence interval instead. This is roughly induced by $\tau_i = 5$. Incorporating conservatism of every agent into the *Standard Model* or, respectively, into the *Sophisticated Model* leads to the two models *Standard-Plus Model* and *Sophisticated-Plus Model*. In the *Standard-Plus Model*, agents behave very similarly to the *Standard Model*, but move only a fraction into the direction of the center, which corresponds to findings on conservatism. The only difference to the *Sophisticated-Plus Model* is simply that we specify the initial signal precision not as equal, but according to the confidence statements. Agents are assumed to know that others are overprecise and thus learn about the original signals by correcting for τ.[21]

Importantly, the four models *Standard Model*, *Sophisticated Model*, *Standard-Plus Model*, and *Sophisticated-Plus Model* are all special cases of Bayesian models and hence produce the

[19] Overprecision, as it is called by Moore and Healy (2008), is also known as "judgmental overconfidence" (Herz et al., 2014), "overconfidence in interval estimates" (Soll and Klayman, 2004), or "resoluteness" (Bolton et al., 2013), and is defined as "excessive certainty regarding the accuracy of one's belief." Conservatism means that agents are not willing to learn sufficiently from new signals (e.g., Peterson and Beach (1967); Mobius et al. (2011); Ambuehl and Li (2018); Mannes and Moore (2013)). Of course, the two patterns are closely related to each other.

[20] Or, alternatively: agents learn from their neighbors, but they attach higher uncertainty to the beliefs of others than to their own belief.

[21] In the conservatism models (consisting of the specifications Standard-Plus and Sophisticated-Plus), we make assumptions about higher-order beliefs that close the model in the sense that no agent will expect another agent to behave in a different manner than in the one observed. In particular, we assume that all agents think of all other agents as overprecise; and that all agents think that all agents think that all agents are overprecise. In that way, an agent i is not surprised that j discounts i's behavior from i's point of view (from a neutral point of view, j takes i's behavior as he should) and that j overvalues j's guess (from i's and a neutral standpoint).

prediction that is formalized as Prediction 1. Except that, in the *Standard-Plus Model* and the *Sophisticated-Plus Model*, agents do not state the same guess x^* from round 3 on, but their subjectively perceived optimal guess x_i^*, which is a mixture between x^* and the agent's initial guess $x_i(1)$. This difference is illustrated in Figure 5 below in the two left panels, which compare the dynamics of the *Standard Model* with the *Standard-Plus Model* in a simple example.

5.2 Specification and Extension of DeGroot models

In the DeGroot framework of naïve learning, agents approach consensus. Consensus is given by $x(\infty) = w'x(1)$, where the vector w captures the eigenvector centrality of the agents (e.g., Friedkin (1991); DeMarzo et al. (2003); Golub and Jackson (2010)).

The most common specification is to allocate equal weights to any connection including to oneself.

$$G = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

Credit for this specification is usually given to DeMarzo et al. (2003). This behavior corresponds to Bayesian updating with independent signals of equal precision in the first round, but not in later rounds. The long-term prediction using this *DeMarzo et al. Model* is determined by $w = (\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})'$, i.e., pendants' initial opinions enter the calculation of the consensus with a weight of 20% each, while the center's initial opinion accounts for 40% of the consensus.

Corazzini et al. (2012) suggest improving the *DeMarzo et al. Model* by increasing the weight of agents who listen to many other agents (and show that this twist improves the model fit to experimental data). The suggested specification is that the weights are proportional to the outdegree (i.e., the number of agents listened to):

$$G = \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} & 0 \\ \frac{3}{4} & 0 & 0 & \frac{1}{4} \end{pmatrix}$$

This model predicts that the center of the star is even more influential in the long run: $w = (\frac{9}{15}, \frac{2}{15}, \frac{2}{15}, \frac{2}{15})'$.[22]

Incorporating conservatism requires a model extension. Friedkin and Johnsen (1990) provide a more general model of naïve learning. Initial opinions are determined by some exogenous

[22] Grimm and Mengel (2016) propose another specification of the DeGroot weights. However, their extension does not lead to an additional prediction here because weights depend on the clustering coefficient, which is zero for all agents in the star network.

conditions, which can always have an impact on an agent's opinion. Such a model has also been analyzed in Golub and Jackson (2012). To incorporate this aspect, we can simply let agents stick to their initial guess $x_i(1)$ to some extent α:

$$x_i(t) = (1 - \alpha_i) \cdot Gx(t-1) + \alpha_i \cdot x_i(1).$$

For $\alpha_i = 0$, we have the DeGroot model. For $\alpha_i = 1$, we have the simplest conceivable model: an agent makes an initial guess $x_i(1)$ and then sticks to it. This is a baseline model that we call the *Sticking Model*, as already mentioned when discussing totally overprecise rational learners.

If $\alpha_i \in (0, 1)$ for every agent i, then the model prediction is that agents move towards the others' guesses, but still rely on their initial guess. This is conservatism.[23] Interestingly, with this model variation, the updating process converges without reaching a consensus (for generic starting values).

We extend the *DeMarzo et al. Model* and the *Corazzini et al. Model* by the Friedkin and Johnsen (1990) framework and set the conservatism/overprecision parameter $\alpha = 0.5$. This leads to the *DeMarzo et al. Plus Model* and the *Corazzini et al. Plus Model*. In these models, agents do not approach consensus anymore. For instance, in the *DeMarzo et al. Plus Model*, the long-term guess of a pendant i is a convex combination of initial guesses with the following weights: weight $\frac{2}{9}$ on the center's initial guess, weight $\frac{1}{27}$ on other pendants' initial guesses each, and weight $\frac{19}{27}$ ($\approx 70\%$) on the own initial guess, which leads to different guesses of each pendant. This difference is illustrated in the right panels of Figure 5. The long-term weights of the *Corazzini et al. Plus Model* are comparable, but differ in that each agent, including the center, is more heavily influenced by the center's initial opinion.

Four models are illustrated in Figure 5. In this example, initial answers are $x_1 = 20\%$ for the center, and $x_2 = 40\%$, $x_3 = 60\%$, and $x_4 = 80\%$ for the pendants. The most important differences are easily observable. In Bayesian models (left panels), learning stops in round 3; in naïve models (right panels), answers converge. In the specifications without conservatism/overprecision (upper panels), agents reach or converge to consensus; in the models with conservatism/overprecision (lower panels), there is a persistent heterogeneity of answers, such that each agent's answer is "biased" towards the own initial answer. Note that the conservative/overprecise agents in the naïve models behave similarly to conservative/overprecise agents in the rational learning approach.

5.3 Comparison of Models (Horse Race)

In total, we have specified nine models. Four following from the rational approach to social learning, four following from the naïve approach to social learning, and one baseline model (the *Sticking Model*), which is a degenerate special case of both model classes. We implemented

[23] Interpretations for the cause of conservatism include forms of overprecision or kinds of anchoring bias in which the initial guess serves as anchor and the adjustments to the others' guesses is limited by parameter α.

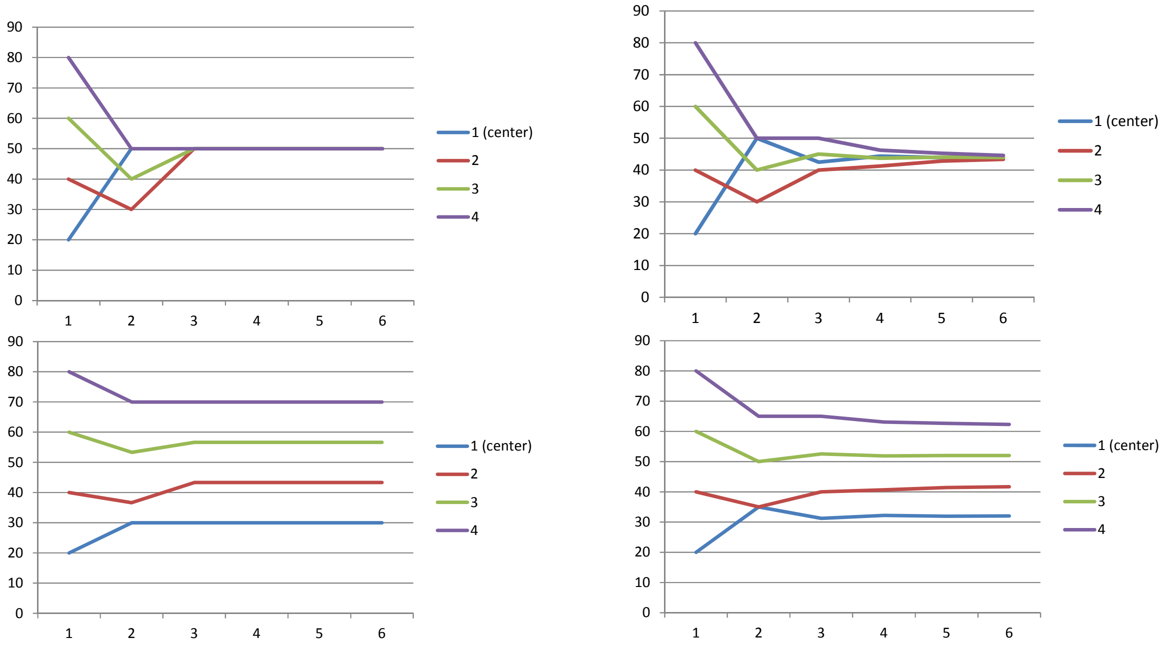

Figure 5: Simple examples of dynamics with time on the x-axis and answers (in percentage points) on the y-axis. Upper panels illustrate two prominent models from the literature; lower panels illustrate their extensions when conservatism is incorporated. *Standard Model* is upper left, *Standard-Plus Model* is lower left, *DeMarzo et al. Model* is upper right, and *DeMarzo et al. Plus Model* is lower right panel. Hence the left panels illustrate rational models, the right panels naïve models.

each model such that all periods $t \geq 2$ are predicted from values at $t = 1$. 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 6 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 "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). Hence, the "Plus" models fit much better than the others in these early periods. However, in the last period, the "Plus" models fit best still, with the only exception that the *DeMarzo et al. Model* fits better than the *Sophisticated-Plus Model.* This observation indicates that the advantage of the models that include conservatism is not restricted to the first rounds, but resists.

Ignoring the "Plus" models for one moment, we can see that the naïve learning in the DeMarzo specification fits well to the data. The sophisticated specification of the rational model does not fit to the data. The standard specification of rational learning and the Corazzini specification of the naïve learning are somewhere in between. Hence, the straightforward specifications that treat all agents symmetrically (*Standard Model*, *DeMarzo et al. Model*) are at least as adequate as the specifications that incorporate confidence statements in a specific way (*Sophisticated Model*), or that incorporate the unequal degree (*Corazzini et al. Model*).

Adding conservatism to the models leads to a very good fit of the rational model in its standard specification (*Standard-Plus Model*) and a better fit of the sophisticated specification (*Sophisticated-Plus Model*) than without conservatism. The best model fit is obtained for the naïve models with conservatism (*Corazzini et al. Plus Model* and *DeMarzo et al. Plus Model*).

We can also differentiate the model fit by treatment. The results are illustrated in Figure A.1 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. Comparisons are similar across treatments. However, the *Corazzini et al. Model* fits better in the accuracy T1 and confidence treatment T2 than in the random treatment T0. The reason is that the center receives a high influence weight in the accuracy and confidence treatment T2, 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.

We finally differentiate between the model fit for the center and for the pendants. The

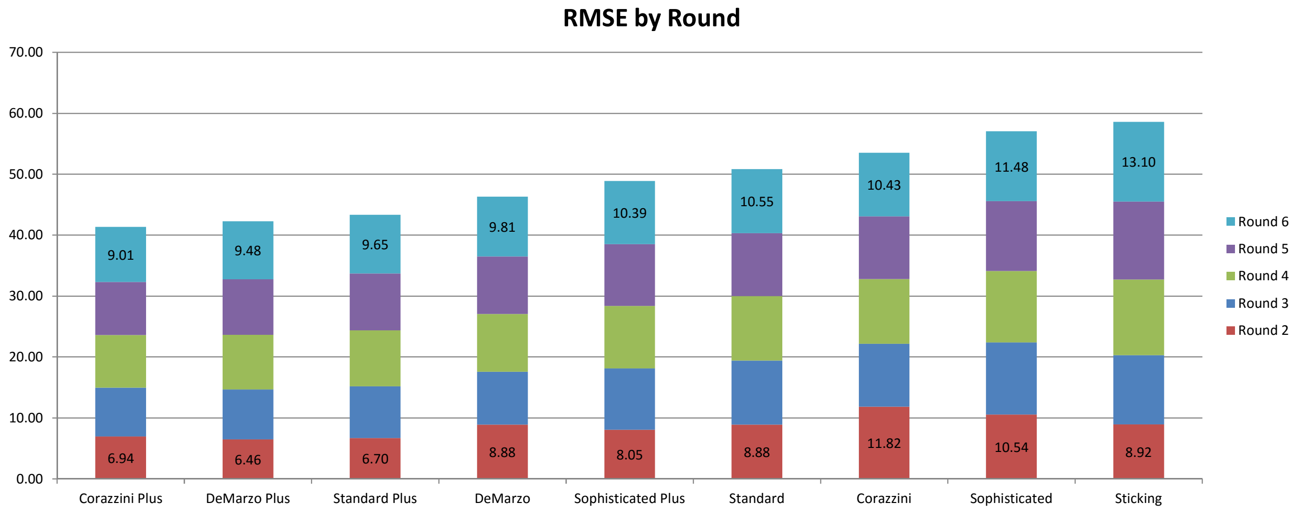

Figure 6: Root mean squared errors (RMSE) of social learning models. "Standard" and "Sophisticated" are models of rational learning; "DeMarzo" and "Corazzini" are models of naïve learning. "Plus" models incorporate conservatism. Lower errors mean better fit between model and data.

result is displayed in Figure A.2 in the Appendix. The *Corazzini et al. Model*, which predicts an immense influence of the center, fits well for the center, but not for the pendants. Again, the "Plus" models fit well generally for both pendants and centers. The best fit for the pendants is attained by the *Corazzini et al. Plus Model*, and the best fit for the center is attained by the *Standard-Plus Model*.

Result 6. *Incorporating "conservatism" into both the rational and naïve models of social learning increases the fit between theoretical models and empirical data.*

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. The exception is that the *Corazzini et al. Model* predicts the center's opinion better than the *Corazzini et al. Plus Model*. Hence, our data strongly indicate 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.

In sum, the results of the horse race show, first of all, that both models of rational and models of naïve learning can contribute to our understanding of social learning in teams. Second, the baseline model that each agent sticks to his own initial guess and keeps his independent opinion fits much better to the data when the team leader was selected at random. Complementarily, models that predict an immense weight of the team leader's opinion (*Corazzini et al. Model*) fit well when the leader is known to be the most confident (T2) or most accurate (T1).

Finally, the known models of social learning might fall short of covering the substantial amount of conservatism that is characteristic for the social learning of real people. Assuming that people are overprecise provides a foundation for conservative learning, even for rational learners, and affects the model prediction such that they are much closer to our data. We can connect this observation with Result 5 that overconfident leaders undermine social learning. Assuming that agents are overprecise induces conservative learning in which the opinions of others are not sufficiently accounted for. Therefore, overconfident leaders undermine performance.

6 Discussion

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 might in principle be helpful. However, 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 leaders' 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. Confidence (T2) is only weakly correlated with actual ability such that the net effect is negative. In addition to a negative effect of declaring the selection criteria of leaders, most people are overconfident in their estimates and in their assessments of problems. Overconfidence may lead to ignorance of the others' valuable opinions, then information gets lost, and the team's potential for solving problems deteriorates. Our data show that indeed overconfidence of both team leaders and other team members have a deteriorating effect.

These are two detrimental effects of leaders selected by confidence. We can further study the micro-mechanisms of these detrimental effects by simulating 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 that 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 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. To introduce this pattern in the theory of social learning, we extend both rational and naïve models by conservatism, which can be derived from overconfidence. With this twist, the fit of each model to the data increases substantially. Moreover, this kind of bounded rationality leads to the fact that leaders learn too little from the opinions in their network.

One conclusion from our paper is that we provide evidence for the advantage of a selection procedure that is based on random leader selection ("sortition"). This process 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. In our setting, selection by confidence leads to detrimental effects on truth-finding, since first the leader listens too little to other members of the group and second the other members listen too much to the leader. Our experiment is one of the first to shed light on one of the potential mechanisms of why aleatory democracy may be beneficial. The strength of random selection is not restricted to reducing the probability of overconfident leaders. It is also based on the fact that the leader's influence on team members is not amplified and, therefore, the others' opinions are respected more compared to a system in which leaders push their own views on all team members due to both a central position in the communication network and an additional legitimacy because they are selected by expertise, or even worse, by confidence.

The problem of overconfidence of leaders and its detrimental effects on group wisdom becomes even more important when considering that expertise is often hard to measure in reality. In fact, publicly expressed subjective confidence in the own expertise might sometimes be more important for becoming a leader than objective expertise. The problem is that the truth is often not precisely known. Therefore, publicly expressed confidence may persuade others that the person in question may know the correct answers. However, as our experiment illustrates, when confidence and expertise are not strongly correlated, overly confident leaders may mislead the group. This difficulty in assessing the true expertise for selecting leaders may therefore be another argument for the beneficial empirical effects of aleatory democracy, where the leaders are selected at random.

In our experiment, we focus on the team's ability to converge to correct assessments of the environment, which is to adapt and learn from each other such that they find correct answers to factual questions. However, it is sometimes less important to find the truth, but more important to converge towards a common opinion. Having a common opinion may help team members to reduce conflicts, to work on the same tasks, and to support each other. This means that opinion convergence can be a separate, distinct goal of social learning in teams and those leaders may be preferable who manage to unify the opinion space in their team. For example, it has been shown that a leader's overconfidence or resoluteness can foster coordination and cohesion (Bolton et al., 2013). In their theoretical contribution, Bolton et al. (2013) already point to the trade-off that an overconfident leader, while having positive effects for coordination, has the downside of not sufficiently learning from the followers. We can now strengthen and empirically document the second mechanism: overconfidence of leaders is clearly a detrimental factor to the team's learning. Hence, the strength of overconfident leaders for coordination comes at the downside of suboptimal information-processing.

6.2 Limitations

The advantages of our experimental design come at the expense of certain limitations. First, 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. In particular, we focus on a guessing task that is assumed to reflect the organizational task of estimating the state of the environment.

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 and less senior employees of the organization. It would also be interesting to see how team members themselves would choose a selection criterion if they were given the opportunity to choose.

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. This is beyond the scope of this paper because it would shift the emphasis from the selection of the team leader to the selection of a communication architecture within an organization. Formal hierarchies within organizations usually have a star-like structure, e.g., they determine the head of an organizational unit, or the president of a certain committee, which can be directly modeled by star networks. However, since informal networks within organizations are also known to be important, alternative network architectures and even endogenous network formation should be considered in future research. The case of the complete network, i.e., every team member can communicate directly with all others, has already been studied (Lorenz et al., 2011) and contrasted to the empty network.

Finding that overconfidence is an important determinant of social learning suggests an alternative treatment that combines accuracy and confidence to overconfidence, in which leaders are selected based on their relatively low or high level of overconfidence. This seems an interesting extension that, however, does not match real selection procedures we are aware of. It could be considered an innovative suggestion to assess overconfidence when selecting managers. Our treatments T1 accuracy and T2 confidence resemble real selection criteria based on maximal competence, which are either objectively assessed (T1 accuracy) or subjectively provided by self-declaration (T2 confidence).

Finally, our experimental design focuses on social learning and does not mix it with the decision-making process. After learning took place in a team, there are various forms of how a decision is actually made. It could be the case that the team communicates its opinions to the higher level management or their client, who then draw their conclusions and take actions. It could also be that the team takes actions on its own, deciding, e.g., by the majority rule or with unanimity about the consequences. Obviously, decision-making processes also affect the quality of the decisions and are thus important to study. However, studying them jointly with the social learning process can distort the measures of learning since communication before collective decisions makes strategic considerations in the communication stage prevalent.

6.3 Practical implications

Despite the limitations of our experiment as discussed above, our findings do suggest several practical implications. First, when selecting a leader, there is a substantial difference between assessing a candidate's competence by some tests (as in our accuracy treatment T1) versus relying on her subjective statement of her own competence (as in our confidence treatment T2). This even holds when there are no strategic incentives to misrepresent the own opinion and the own competence. Our findings clearly suggest, whenever possible, focusing on objective measures of competence rather than trusting subjectively stated confidence in candidates' own expertise. A large majority of people is overconfident, such that starting a competition as to who is claiming the highest confidence will most likely lead to detrimental effects in selecting leaders who will listen too little to other opinions in their network. Hence, when the main goal of the team is related to truth-finding, this is expected to be a poor selection criterion.

Second, the way the selection criterion for the leader is communicated to a team heavily affects the team's interaction and performance. In particular, making explicit that the team leader was selected at random can lead other team members to make use of their own valuable knowledge instead of "blindly" following their leader. By the hierarchical structure, which determines the communication network, the team leader is already very powerful and her opinion is certainly heard. Declaring that the team leader was selected because of her (alleged) superiority increases her power, which might push team learning out of balance. Hence, keeping quiet about the (alleged) superiority of a team leader can foster more efficient learning within a team.

Third, we can validate that communication and social influence can be harmful for the wisdom of crowds effect (Lorenz et al., 2011). We confirm this finding for unequal communication structures in terms of star networks, by showing that the wisdom of crowd error increases over time. This provides evidence that the group can exploit less and less information from other network members over consecutive periods of social influence. 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. 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, but prevent single 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.

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. 2016. “An Experiment on Belief Formation in Networks.” *Available at SSRN 2361007* .

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.

Mobius, 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

	(1) Exp. Payoff (EUR)	(2) Exp. Payoff (EUR)	(3) Exp. Payoff (EUR)
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 (average)			-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)	(2)	(3)
	T0 random: answer_6	T1 accuracy: answer_6	T2 confidence: answer_6
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.

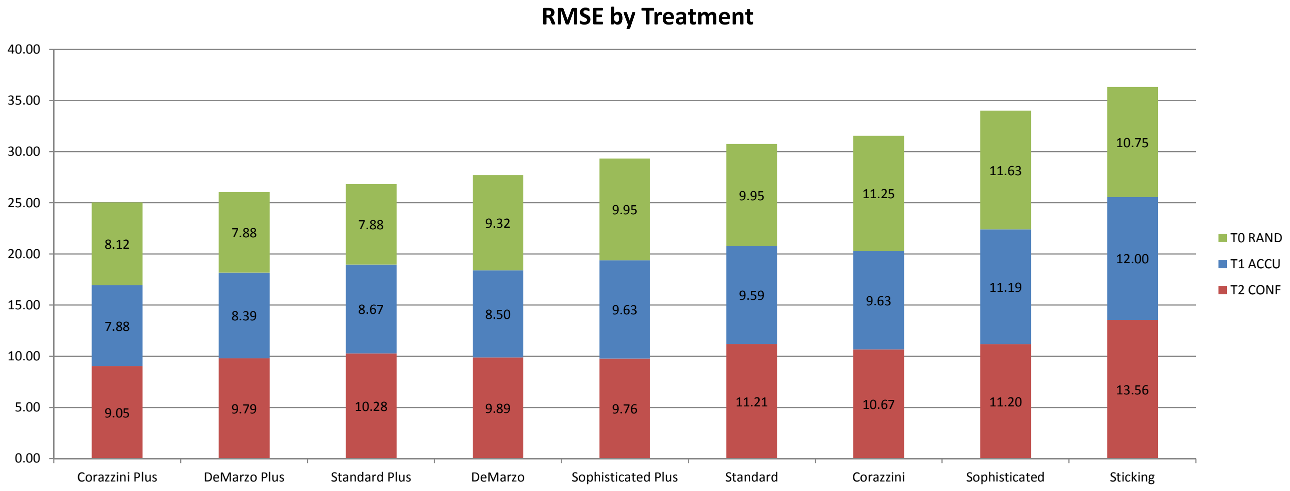

Figure A.1: Root mean squared errors (RMSE) of social learning models differentiated by treatment. Lower errors mean better fit between model and data.

	(1) T0 random: answer_6	(2) T1 accuracy: answer_6	(3) T2 confidence: answer_6
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.

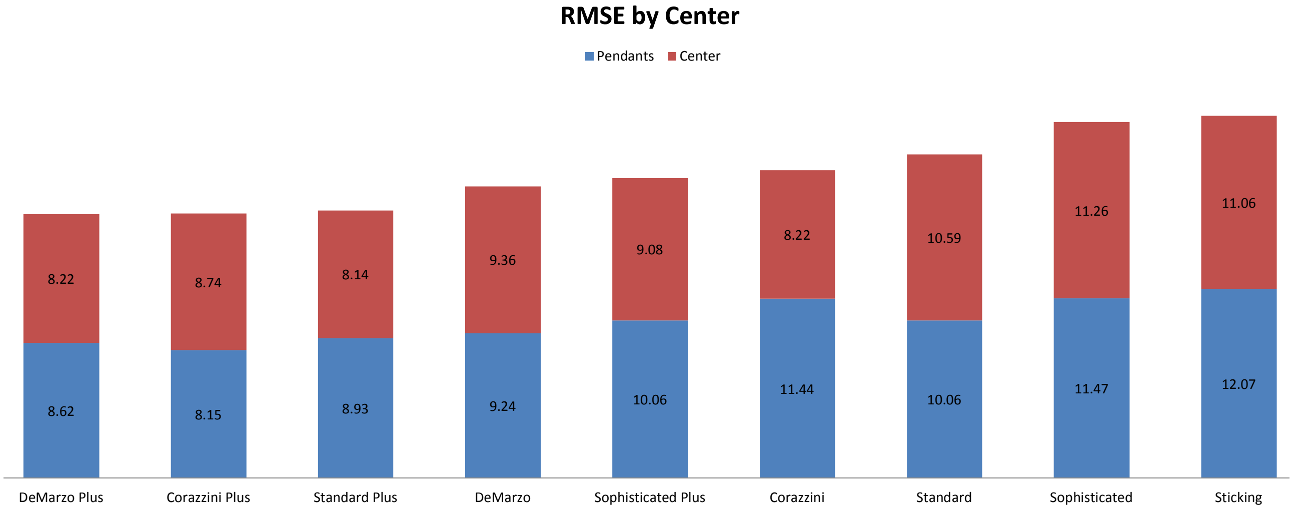

Figure A.2: Root mean squared errors (RMSE) of different models by center and pendants differentiated by center and pendants. Lower errors mean better fit between model and data.

	(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 (average)	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 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 correpsonding 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.