The Biases of Others: Projection Equilibrium in an Agency Setting*

David N. Danz[†], Kristóf Madarász[‡], Stephanie W. Wang[§]

This version: October 31, 2018.

First Online Draft available since August 2014.

Abstract

We study the structure of biased social cognition which involves not simply one's belief about the beliefs of others, but also one's belief about their beliefs of one's own belief. We find that while people naively project their information onto differentially-informed others, they also anticipate differentially-informed others' projection onto them. In a principal-agent setting, we directly test the tight one-to-one structural relationship between the partial extent to which the typical person projects her information onto others, ρ, and the extent to which she anticipates but partially underestimates the projection of others onto her, ρ^2. The data is remarkably consistent with the parsimonious link implied by the model of projection equilibrium. Furthermore, the majority of subjects both think that others are partially biased, but they also partially underestimate the extent of their bias. The result lends support to the notion of biased social cognition arising as a combination of a biased, but coherent fully ego-centric belief anchor with a partial probabilistic adjustment to the truth.

Keywords: structure of biased beliefs, projection, theory of mind, partial adjustment, behavioral organizations.

Keywords: social cognition, theory of mind, biased higher-order beliefs, projection equilibrium, behavioral organizational economics.

*We are grateful to Ned Augenblick, Douglas Bernheim, Arie Beresteanu, Colin Camerer, Gary Charness, John Duffy, Ignacio Esponda, Dorothea Kübler, Muriel Niederle, Demian Pouzo, Al Roth, Andrew Schotter, Adam Szeidl, Lise Vesterlund, Axel Werwatz, Alistair Wilson, and seminar audiences at Berkeley, Columbia, CMU, University College London, University of Southern California, Stanford Econ, Stanford SITE 2014, and Utah for comments. Financial support from the Deutsche Forschungsgemeinschaft (DFG) through CRC 649 "Economic Risk" is gratefully acknowledged. First Online Draft: https://site.stanford.edu/2014/session-7

[†]University of Pittsburgh and WZB Berlin.

[‡]London School of Economics and Political Science and CEPR

[§]University of Pittsburgh.

"I found the concept of hindsight bias fascinating, and incredibly important to management. One of the toughest problems a CEO faces is convincing managers that they should take on risky projects if the expected gains are high enough. [...] Hindsight bias greatly exacerbates this problem, because the CEO will wrongly think that whatever was the cause of the failure, it should have been anticipated in advance. And, with the benefit of hindsight, he always knew this project was a poor risk. What makes the bias particularly pernicious is that we all recognize this bias in others but not in ourselves."
R. Thaler, *Misbehaving* (2015).

1 Introduction

Despite growing interest in understanding how people's beliefs actually deviate from the truth (e.g., Tversky and Kahneman 1974, Camerer et al. 2004, Gennaioli and Shleifer 2010, Bénabou and Tirole 2016, Augenblick and Rabin 2018, Jehiel, 2018), there is much less careful evidence on the metacogniton about such tendencies, in particular, on whether people anticipate the biases of others. Do they explicitly think that others form biased beliefs and if so, how well calibrated they are? What is the relationship between people's awareness of the biases of others and the presence of the same kind of bias in their own judgement? These issues have direct economic consequences and studying them can also guide the kind of approaches aimed at increasing the realism of economic models of social behavior.

The structure of an individual bias, such as a person's misprediction of her future preferences, can be described independently from her anticipation of the same kind of mistake in others.[1] The same is not true for a social bias, that is, when people have biased beliefs about the beliefs of others. An agent's perception of how her principal thinks entails how he thinks she thinks. In turn, the definition

[1] In self-control problems, evidence suggests that people are too optimistic when predicting the time they will take to complete a task, but more pessimistic when predicting the time others will take, Buehler, Griffin, and Ross (1994). See also Pronin et al. (2002) in the context of people thinking that others are more overconfident than they themselves are. In the context of loss aversion, van Boeven et al. (2000) provide evidence consistent with the idea that people anticipate the fact that ownership changes preferences but underestimate the extent of this change both for themselves and for others.

of a social bias must specify a person's mistake and her perception of the mistakes of others simultaneously.

Our paper then considers the domain of thinking about how others think, forming beliefs about their beliefs. This domain, sometimes referred to as theory of mind capacity, is essential for social cognition and strategic behavior. Furthermore, in the presence of private information, evidence from 'false-belief tasks' (Piaget, 1953; Wimmer and Perner, 1983), hindsight tasks (Fischhoff, 1975), curse-of-knowledge tasks in markets (Camerer et al., 1989), the illusion of transparency (Gilovich et al., 1998), or the outcome bias (Baron and Hershey, 1988), point to a robust shortcoming in this domain.[2] People systematically under-appreciate informational differences in that, they too often act as if others had the same information they did.

Strategic choice, however, depends not simply on what basic information a person assigns to others, i.e., her estimate of her opponent's first-order beliefs about the payoff state, that matters. Instead what a player thinks others think she knows, and so on, is often equally key. To even formulate the phenomenon of such informational projection, one must consider a person's basic mistake and her model of how others think, which includes her perception of this tendency in others, simultaneously. At the same time, we are unaware of any prior empirical or experimental evidence that carefully documents or structurally evaluates the joint presence of such beliefs.

Anticipating the biases of others has direct economic implications. In agency settings a principal—an investor or a board—evaluates the quality of an agent—an executive, a doctor, or a public bureaucrat—by monitoring him with ex-post information. A principal who naively projects such ex post information exaggerates how much the agent should have known ex ante. In turn, she misattributes the informational gap between the ex ante and ex post stages to the lack of sufficient skill (or ex ante effort) by the agent and underestimates the agent's quality on average. An agent who anticipates that the principal is biased will then want to 'cover his ass' and engage in defensive practices aimed at reducing this gap, e.g., distort the production of ex-ante information, undertake an ineffective selection

[2] Although hindsight bias is sometimes described as an intrapersonal phenomenon, the evidence is predominantly from interpersonal settings.

of tasks, or simply dis-invest from an otherwise efficient relationship, (Madarasz, 2012).[3] Crucially, the presence of such defensive agency depends then *jointly* on the principal naively exhibiting the basic bias *and* the agent also anticipating that the principal has this biased tendency.

As an other example, consider classic common-value trade between an informed dealer and an uninformed buyer as in Akerlof (1970). Suppose that the naive buyer projects her ignorance and thinks that the dealer also does not know whether the car is a peach or a lemon. It is an informed dealer who anticipates the buyer's projection, and, in turn, that the buyer under-appreciates the presence of the winner's curse, who will be prompted not to disclose the product's quality and still quote an excessive price whenever selling what she knows to be a lemon. Systematic deception is a function of *both* the buyer projecting her ignorance on the dealer, thinking that he does not know what she does not know, and the dealer anticipating that the buyer has biased beliefs about what the dealer knows, anticipating that her beliefs of his beliefs are too far from his beliefs.

Given the robust presence of the basic mistake, one may argue that its anticipation in others shall be uncommon; being aware of this mistake in others should prompt a person to recognize and correct this tendency in herself. A salient approach may then be to adopt a dichotomous classification; a person is either naive in that she is subject to this mistake and does not anticipate it in others, or she does not exhibit this mistake and is (at least partially) sophisticated about the presence of this mistake in others. Under this classic dichotomy, which is then in contrast to what is suggested by the quote from Thaler (2015) above, the implications of this phenomenon depend directly on the way these different types sort into different roles. An organization may best response by adopting an effective sorting method of navies and sophisticates. The test of whether such a sharp di-

[3] A widely discussed example of such defensive agency can be found in the context of medical malpractice liability. The radiologist Leonard Berlin, in his 2003 testimony on the regulation of mammography to the U.S. Senate Committee on Health, Education, Labor, and Pensions, describes explicitly how ex-post information causes the public to misperceive the ex-ante visual accuracy of the mammograms produced by radiologists, implying that juries are "all too ready to grant great compensation." Berlin references the role of information projection in such ex-post assessments, where ex-post information makes reading old radiographs much easier. In response, physicians are reluctant to follow such crucial diagnostic practices: "The end result is that more and more radiologists are refusing to perform mammography [...] In turn, mammography facilities are closing."

chotomy provides a good approximation of the perceptual heterogeneity in reality is then critical, but also appears to be missing from the literature.

Moving beyond this classic dichotomy, there is a bewildering variety of ways in which a comprehensive account of the higher-order implications of this phenomenon may be specified. At the other end of the above dichotomy, one could opt for a fully flexible approach, with potentially many degrees of freedom describing such implications, in effect rejecting the idea of a parsimonious relationship between the basic mistake and the anticipation of this mistake in others.

In contrast to such a skeptical stance, the model of projection equilibrium, Madarasz (2014, revised 2016), offers a fully specified yet parsimonious general account. It proposes a model of partial projection governed by a single scalar $\rho \in [0, 1)$. It postulates that a person's belief hierarchy is a probabilistic mixture of an all-encompassing projective fantasy of her opponent and an unbiased view of her opponent. This projective fantasy is all-encompassing in the sense that a person mistakenly believes that with probability ρ her opponent not only has access to the same basic information about the payoff state as she does, but also knows the way she thinks, i.e., her entire belief hierarchy. A player then partially adjusts her expectations to the truth by placing the remaining weight on an unbiased estimate of how her opponent thinks. The model ties together the extent of one's basic mistake and her anticipation of the mistake of others and pins down the structure of mispredictions along each player's belief hierarchy.

The model implies a one-to-one relationship between the partial extent to which a player projects onto others (*first-degree projection*) and the extent to which she anticipates, but partially underestimates the projection of others onto her (*second-degree projection*). Our paper then introduces an experimental design to understand the key issue of anticipation and the extent to which this account may help organize key aspects of the data.

In our experiment, principals estimated the average success rate π of reference agents in a real-effort task. While agents never received the solution to the task, principals received the solution to the task prior to the estimation in the asymmetric informed treatment, as in the case of monitoring with ex-post information, but not in the symmetric uninformed treatment. Projection equilibrium predicts that a principal in the former, but not in the latter treatment, should systematically overestimate the agents' success rate on average. We find that in the uninformed treatment principal are well-calibrated. In contrast, consistent with previous results, in the asymmetric treatment find a very strong exaggeration. While the true success rate is 39% the principals' estimate is 57% on average. This difference allows us to identify the extent of first-degree projection in our data.

To first obtain a qualitative response regarding anticipation, agents could choose between a sure payoff and an investment whose payoff was decreasing in the principal's estimate of the success rate of the other agents performing the task. Consistent with the comparative static prediction of projection equilibrium, we find strong evidence of anticipation of projection in that 67.3% of agents in the informed treatment as opposed to 39.2% in the uninformed treatment chose the sure payment over the investment whose payoff was decreasing in the principal's estimate. In the context of defensive agency, the fraction of instances where the agent, a doctor, a manager or an administrator is willing to take on ex ante risk when she is monitored with ex post information drops by thirty percent.

We then turn to the main point of our paper. We elicited both agents' first-order and second-order estimates (their estimate of the principals' estimates) of the success rate of the reference agents. In the symmetric treatment, projection equilibrium, just as the unbiased BNE, predicts that an agent's first- and second-order estimates should be unbiased *on average* and also equal the the principal's first-order estimate. Indeed the data confirms all of these predictions. In contrast, in the informed treatment, while the same equivalence must hold under the unbiased BNE, projection equilibrium predicts two key departures. Specifically, while the agent's first-order estimate shall be correct on average, (i) her second-order estimate shall be higher than her own first-order estimate and (ii) her second-order estimate shall be lower than the principal's first-order estimate on average. We find exactly this pattern. The agents' second-order estimate is 51%. As implied by projection equilibrium players explicitly anticipate that others are biased but underestimate its extent.

Next we consider the link between the first- and second-degree projections. The model predicts that the extent to which the principal exaggerates the success rate in the informed treatment, shall fully pin down the extent to which the agent under-estimates the principal's exaggeration in this treatment on average. If the former is $\rho(1-\pi)$, the latter shall be $-\rho^2(1-\pi)$ and vice versa. In our data the degree of projection calculated based on the principal's exaggeration of the success rate is 0.326 while that calculated based on the agent's underestimation of this exaggeration is 0.334. We then perform a more careful econometric estimation and show that the estimate which allows these two to freely differ and the one which constraint these to be the same delivers very similar parameter estimates and very similar log-likelihoods. The data thus directly supports the logic of projection equilibrium.

Finally, we also describe (i) the distribution of the degree of projection inferred from the principal population and (ii) the distribution of the degree projection inferred from the agent population. We document three main facts. First, the majority of the principals are partially biased, they partially exaggerate the success-rate of the agents. Second, the majority of the agents also exhibit a partial bias; they believe that principals are partially biased but underestimate its extent. Indeed, for the majority of our agents we can reject both the hypothesis that they are fully naive about the biases of others *and* also that they fully anticipate the biases of others. Instead we find that the majority of people believe that others are *partially* biased and also partially underestimate the extent to which they are. Third, and perhaps most surprisingly, we find that these two distributions are remarkably close to each other. In sum, the data rejects the classification of people into naive and sophisticated types and instead provides strong support for the model of projection equilibrium and, more generally, to the underlying idea of a social belief bias arising from a fully-egocentric but logically coherent anchor with a partial adjustment to the truth.

To the best of our knowledge, our paper is the first to consider a structured test of people's beliefs about other people's biased beliefs while also demonstrating the impact of these on strategic behavior. The rest of the paper is organized as follows. Section 2 presents the design, Section 3 the predictions, Section 4 the results. In Section 5 we discuss alternative hypotheses and the issue of conditional beliefs which provide further support for the mechanism proposed. In Section 6 we conclude.

2 Experimental Design

2.1 Experimental task

All participants worked on the same series of 20 different change-detection tasks. In each basic task the subjects had to find the difference between two otherwise identical images. Figure 1 shows an example. The change-detection task is a common visual stimulus (Rensink et al., 1997; Simons and Levin, 1997) and has already been studied in the context of the curse-of-knowledge, Loewenstein et al. (2006).

Figure 1: Example of an image pair. Image sequence in the experiment: A, B, A, B, ...

We presented each basic task in a short clip in which the two images were displayed alternately with short interruptions.[4] Afterwards, subjects could submit an answer by indicating the location of the difference on a grid (see Instructions in the Appendix).[5]

2.2 Principals

Principals had to estimate the performance of others in each basic tasks. Specifically, the principals were told that subjects in previous sessions worked on the tasks and that these subjects (*reference agents*, henceforth) were paid according

[4] Each image was displayed for one second, followed by a blank screen for 150 milliseconds. The total duration of each clip was 14 seconds.

[5] See the instructions in the Appendix for more details.

to their performance. The performance data of 144 reference agents was taken from Danz (2014) where the subjects performed the tasks in winner-take-all tournaments and where they faced the tasks in the exact same way as the subjects in the current experiment.

In each of the 20 rounds, the principals were first exposed to a new basic task. Afterwards, the principal stated her estimate (b_P) of the fraction of reference agents who spotted the difference in that task (success rate π henceforth). After each principal stated his or her belief, the next round with a different basic task started.[6]

For the principals the two treatments differed as follows. In the *informed* (assymetric) *treatment*, principals received the solution to each basic task before they went through the change-detection task. Specifically, during a countdown phase that announced the start of each task, the screen showed one of the two images with the difference highlighted with a red circle (see Figure 2). This mimics various motivating economic examples, e.g., monitoring with ex post information after an accident, a realized portfolio allocation decision, or a medical outcome where the principal learns the ex post outcome and the case solved by the agent ex ante at the same time. In the *uninformed* (symmetric) *treatment* instead, the principals were not given solutions to the basic tasks (the same image was shown on the countdown screen, but without the red circle and the corresponding note in Figure 2). Principals in both treatments then went through each task exactly as the reference agents did. The principals did not receive any feedback during the experiment.

At the end of the sessions, the principals received €0.50 for each correct answer in the uninformed treatment and €0.30 in the informed treatment. In addition, they were paid based on the accuracy of their stated estimates in two of the 20 tasks (randomly chosen): for each of these two tasks, they received €12 if $b \in [\pi - 0.05, \pi + 0.05]$, that is, if the estimate was within 5 percentage points of the true success rate of the agents. We ran one session with informed principals, and one with uninformed principals with 24 participants in each.

[6] The principals first participated in three practice rounds to become familiar with the interface.

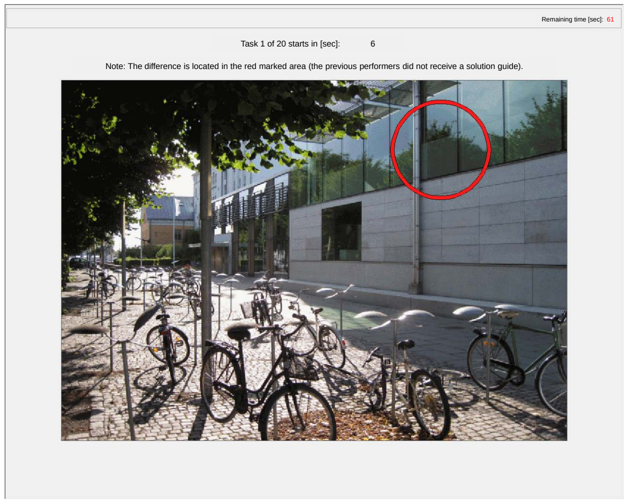


Figure 2: Screenshot from the treatment with informed principals: Countdown to the next task providing the solution (translated from German).

2.3 Agents

Agents in our experiment, in both treatments, were informed that in previous sessions reference agents had performed the basic tasks being paid according to their performance on the basic task and that principals had estimated the average performance of these reference agents being paid according to the accuracy of their estimates. The agents were further told that they had been randomly matched to one of the principals at the outset of the experiment and that this matching would remain the same for the duration of the experiment.

For the agents, the two treatments differed solely with respect to the kind of principal they were matched to: in the informed treatment, agents were randomly matched to one of the informed principals; in the uninformed treatment, agents were randomly matched to one of the uninformed principals. In both treatments the agents where made fully aware whether or not the principal has received the solution (both of course not of the existence of the other treatment).

In each of the 20 rounds, the agents in both treatments first performed the basic task in the same way as the reference agents did; that is, they went through the images and then submitted a solution. Afterwards, following each of the first 10 change-detection tasks, the agent, in both treatments, stated his estimate of the fraction of reference agents who spotted the difference in that task (*first-order estimate* b_A^I henceforth) and his estimate of the principal's estimate of that success rate (*second-order estimate* b_A^{II} henceforth). For the second 10 change-detection tasks, the agent in both treatments decided between two options A and B. Option A provided a sure payoff of €4. Option B was a lottery where the agent received €10 if the principal's estimate b_P was not more than 10 percentage points higher than the success rate π; otherwise, the agent received €0. This decision, implicitly, is also a function of the agent's first and second-order beliefs about the success-rate. Choosing option B can be thought of as an investment whose perceived expected return is decreasing in the wedge between the agent's second- and first-order estimates. Throughout the paper, we will refer to this choice as the agents' investment decision.

We have also ran separate sessions without belief elicitation, that is, where the agents, following each of the 20 change-detection tasks, after solving this task had to choose between option A and option B as described above. In the result section we also present this data.

Agents also received feedback, in the exact same way in both treatments, regarding the solution to each task right after solving the task. Specifically, the screen showed one of the two images with the difference highlighted with a red circle; then, the images were shown again. Agents matched to informed principals were told that this feedback corresponded to what the principal had seen for that task. Agents matched to uninformed principals were told the principals had not received this solution to the task. In neither treatments, did agents receive any information about the principals' estimates.

Finally, agents received €0.50 for each correct answer to the change-detection tasks. In addition, at the end of the experiment one round was randomly selected for additional payment. If this round involved belief elicitation, we randomly selected one of the agent's stated estimate for payment, namely, either her first- or second-order estimate in that round.[7] The subject received €12 if her stated estimate was within five percentage points of the actual value (the actual success rate in case of a first-order estimate and the principal's estimate of that success rate

[7]This payment structure addresses hedging concerns (Blanco et al., 2010).

in case of a second-order estimate), and nothing otherwise.[8] If the round selected for payment involved an investment decision, the agent was paid according to her decision.

2.4 Procedures

The experimental sessions were run at the Technische Universität Berlin in 2014. Subjects were recruited with ORSEE (Greiner, 2004). The experiment was programmed and conducted with z-Tree (Fischbacher, 2007). The average duration of the principals' sessions was 67 minutes; the average earning was €15.15. The agents' sessions lasted 1 hour and 45 minutes on average; the average payoff was €20.28.[9] Participants received printed instructions that were also read out loud, and had to answer a series of comprehension questions before they were allowed to begin the experiment.[10] At the end of the experiment but before receiving any feedback, the participants completed the four-question DOSE risk attitude assessment (Wang et al., 2010), a demographics questionnaire, the abbreviated Big-Five inventory (Rammstedt and John, 2007), and personality survey (Davis, 1983).

3 Predictions

The predictions below are based directly on the model of projection equilibrium, Madarasz (2016).[11] To describe these for our design, let there be a set of payoff states Ω, with generic element ω, and a prior ϕ over it. Player i's information about the state is generated by an information partition $P_i : \Omega \to 2^{\Omega}$, her action set is given by A_i, and her payoff function by $u_i(\omega, a)$ where a is an action profile. The game is then summarized by $\Gamma = \{\Omega, \phi, P_i, A_i, u_i\}$.

[8] We chose this elicitation mechanism because of its simplicity and strong incentives. In comparison, the quadratic scoring rule is relatively flat incentive-wise over a range of beliefs, and its incentive compatibility is dependent on assumptions about risk preferences (Schotter and Trevino, 2014). The Becker-DeGroot-Marschak mechanism can be confusing and misperceived (Cason and Plott, 2014). The beliefs we elicited were coherent and sensible.

[9] The average duration of the sessions (the average payoff) in the treatments with and without belief elicitation was 115 and 96 minutes (€21.47 and €19.10), respectively.

[10] Two participants did not complete the comprehension questions and were excluded from the experiment.

[11] See https://works.bepress.com/kristof_madarasz/43/.

Solving the basic task amounts to picking a cell $x \in D$ from the finite grid on the visual image. This is performed by all players. In the game corresponding to our design, there are only two strategically active players: one agent and one principal. The reference agents are strategically passive; they perform only the basic task. Furthermore, they have a dominant strategy of maximizing the probability of success. In what follows then subscript A refers to the strategically active agent and subscript P to the principal. The action set of the principal, which includes his estimation task, is $A_P = D \times [0,1]$. The action set of the strategically active agent, in the rounds where his two estimates are elicited, is $A_A = D \times [0,1] \times [0,1]$. Since for no player i does the payoff from choosing x_i directly interact with the payoff from any other decision, we denote this payoff by $f(x_i, \omega)$ and normalize it to be one if the solution is a success and zero otherwise.[12]

Projection Equilibrium. Under projection equilibrium each player i, Annie, best responds to a bias perception of her opponent j's, Paul's, strategy. Annie assigns probability ρ_i to a strategy played by a fictional projected version of Paul and probability $1 - \rho_i$ to Paul's true strategy. The projected version of Paul conditions his strategy on the exact same information about the payoff state as Annie does, thus Annie projects both her basic information and her ignorance, *and* best responds to Annie's true strategy.[13]

Under projection equilibrium, each player i acts on the basis of a coherent but biased belief hierarchy implicit in the above heuristic definition. This belief hierarchy, which directly determines the predictions of the model for our design, is determined by two key and simple psychological features. First, projection is all-encompassing. Annie, assigns probability ρ_i to the projected version of Paul. This fictional version of Paul shares, in each state, not only Annie's information about the payoff state; he also assigns probability 1 to Annie's actual belief hierarchy. He knows Annie's beliefs. In the context of our design, Annie thinks that the projected version of Paul has the same beliefs about the solution to the basic

[12]In our design, strategically active players always estimate the success rate of the strategically passive agents. This ensures that there cannot be an equilibrium where the agent and the principal may co-ordinate on sub-optimal performance on the basic task to achieve a higher earning on the estimation tasks.

[13]In an N-person game, the projected versions of player i's respective $N-1$ opponents occur in a perfectly correlated manner, and player i believes that these projected versions of her opponents know this. For a detailed presentation and discussion, see Madarasz (2016).

task as she does, that Paul knows exactly what Annie believes about the solution to this basic task, knows exactly her belief about what others believe about the solution to the basic task, and so on. Second, Annie assigns positive probability $1-\rho_i$ to the real version of Paul, that is, an, on average, unbiased estimate to how he thinks and behaves in reality.

Bias Structure. In sum, Annie's expectations are anchored to a coherent all-encompassing projective fantasy of Paul which is adjusted partially and probabilistically to the truth. The weight $1-\rho_i$ measures Annie's partial adjustment to reality away from her fantasy. Projection equilibrium, via the above two assumptions of all-encompassing projection and partial adjustment to the truth, pins down the meaning of informational projection for strategic settings. It implies a polynomially vanishing tight bias structure along each player's belief hierarchy. In particular, it postulates a parsimonious one-to-one relationship between the extent to which a player projects onto her opponent and the extent to which she anticipates but under-appreciate the projection of her opponent onto her. The predictions below will highlight this relationship.

Heterogeneous Projection. We first state the predictions under heterogeneous role-specific projections, that is, we allow the principal and the agent to project to differing degrees, i.e., $\rho_P \neq \rho_A$ may hold. This specification will already greatly restrict the set of possible outcomes in our design.

Homogeneous projection. We then state the predictions under homogeneous projection, $\rho_A = \rho_P$. This case is of particular interest in our design. Since we infer the basic bias from the choice of the principal and the misperception of the principal's bias from the choice of the agent, the homogeneous case allows us to directly test the tight link which postulates that the former should fully determine the latter and vice versa.

Before turning to the predictions some additional remarks are in order. Below, we do not assume that people make the same inference from watching the video of the alternating images per se. Instead, we allow players to obtain different private signal realizations from watching the video, that is, about the change-detection task. We assume only that, from the relevant ex-ante perspective, that is, before the identity of each player is randomly determined, the distribution of these signal realizations is the same for each player. In turn, the predictions below hold in the ex-ante expected sense. We focus on the average estimate within each treatment and the predictions below express differences about such average estimates across treatments. Later in the paper we return to the link between conditional estimates and projection. Finally, the predictions below nest the unbiased BNE, the prediction based on, on average, unbiased beliefs, i.e, $\rho_i = 0$ for $i \in \{A, P\}$.

Consider first the ex-ante probability with which a randomly chosen player i who sees only the video can solve the basic task. Denote this success rate by π. Formally, let

$$\pi \equiv E_\omega[\max_{x \in D} E[f(x,\omega) \mid P_A(\omega)]].$$

Consider now the ex-ante expected difference between the above probability and the success probability on the basic task by the principal. Formally, let

$$d \equiv E_\omega[\max_{x \in D} E[f(x,\omega) \mid P_P(\omega)]] - E_\omega[\max_{x \in D} E[f(x,\omega) \mid P_A(\omega)]].$$

In the uninformed (symmetric) treatment neither the agent nor the principal is given the solution. Hence, by the law of iterated expectations, $d = 0$ must hold. In the informed (asymmetric) treatment, the principal also has access to the solution. Since the solution always helps solve the change-detection task, $d > 0$ must hold.

We can now turn to the predictions. The predictions already follow from interim iterative dominance, e.g., Fudenberg and Tirole (1983), given the structure of biased beliefs implied by the model. Hence, they not rely on a fixed-point argument.[14]

Claim 1. *Under projection equilibrium the principal's ex-ante expected estimate of* π *(denoted by* b_P^I*) is* $\pi + \rho_P d$.

In the uninformed treatment the principal's estimate is unbiased. Her estimate *conditional* on her own success or failure on the basic task may well be affected by projection, e.g., it may be inflated following own success and deflated following own failure. Such distortions, which cannot be pinned down in the absence of

[14]The predictions below also rest on the assumption that people report the mean of their expectations.

further assumptions, however, must cancel out on average. This follows from the fact that the agents and the principal have the same ex-ante probability of success on the basic task in this treatment and roles are determined randomly. Even if the principal fully projects; predicts success whenever she figures out the solution and likely failure whenever she does not, her estimate is correct *on average*.

In the informed treatment, in contrast, the principal always knows the solution, hence does so more often than the agents. A projecting principal then exaggerates the probability with which the reference agents shall figure out the solution on average, and does so by $\rho_P d$.

Claim 2. *Under projection equilibrium the agent's ex-ante expected*

1. *estimate of* π *(first-order estimate* b_A^I*) is* π*;*
2. *estimate of the principal's estimate of* π *(second-order estimate* b_A^{II}*) is* $\pi + (1-\rho_A)\rho_P d$.

The model implies a systematic wedge between an agent's *own* second-order and first-order estimates in the informed but not in the uninformed treatment. The agent's first-order estimate is predicted to be unbiased in both treatments. The reason is the same as for the principal in the uninformed treatment described above. In the uninformed treatment the same holds for all estimates. Just as under the unbiased belief hierarchy supporting the unbiased BNE, all estimates are predicted to be equal to π on average. In short, on average, $b_P^I = b_A^{II} = b_A^I = \pi$ must hold for any $\rho_A, \rho_P \geq 0$.

In the informed treatment, the same equivalence holds under the unbiased BNE. Projection equilibrium instead predicts two departures. First, the agent's second-order estimate is predicted to be systematically *higher* than his *own* first-order estimate. Second, his second-order estimate is also predicted to be systematically *lower* than the principal's first-order estimate. The former is due to the fact that the agent anticipates the principal's projection. The latter is due to the fact that, in proportion to his own projection onto the principal, he underestimates the principal's exaggeration. In short, on average, $b_P^I > b_A^{II} > b_A^I = \pi$ holds iff $\rho_A, \rho_P > 0$.

To describe the logic, note that if $\rho_A = 0$, only the second inequality is strict. An *unbiased* agent does not project and fully anticipates the principal's bias, hence, there is no wedge between his second-order estimate and the principal's first-order estimate. In contrast, if $\rho_A \to 1$, only the first inequality is strict and $b_A^{II} = b_A^I$. A *fully* biased agent thinks that the principal (and the reference agents) always believes the same thing about the solution to the task as he does and, by virtue of projection being all-encompassing, that the principal always knows what he (and the reference agents) believe about the solution to the task. In turn, there is no wedge between the agent's first-order and second-order estimates. Given a *partially* biased agent, however, under projection equilibrium, all of the above inequalities hold strictly.

The implication of the model to our design under heterogeneous projection is that the agent anticipates but partially underestimates the principal's exaggeration of π in the informed treatment. Under homogeneous projection the predictions are further refined. Here, by virtue of all-encompassing projection, the extent of the principal's exaggeration of the success rate, driven by her own projection onto the agents, fully pins down the extent to which the agent underestimates this exaggeration, due to the agent's own projection onto the principal, on average and vice-versa.

The table below summarizes the predicton of the polynomially vanishing bias structure along the player's belief hierarchy in this case.

Ex-ante expected bias	Uninfo Treatment	Info Treatment
principal's first-order estimate	0	$\rho(1-\pi)$
agent's first-order estimate	0	0
agent's second-order estimate	0	$-\rho^2(1-\pi)$

Finally, as mentioned, we also consider a less structured setting where the principal's action set is unchanged, but where the strategically active agent's action set is $A_A = D \times \{\text{Invest, Not Invest}\}$.

Claim 3. *Under projection equilibrium, iff $\rho_A, \rho_P > 0$, the agent's propensity to invest is lower in the informed than in the uninformed treatment on average.*

The model implies that the agent attaches weight ρ_A to the projected version of the principal and weight $1 - \rho_A$ to an unbiased estimate of the real version of the principal. The agent's estimate of the projected version's beliefs, conditional on any given performance of the agent on the basic task, is the exact same in both treatments. The agent's estimate of the real version's beliefs, again conditionally on any given own performance on the task, is strictly higher, in the sense of first-order stochastic dominance, in the informed than in the uninformed treatment. Hence, the agent is predicted to invest less often in the lottery whose return is decreasing, in the sense of first-order stochastic dominance, in the principal's beliefs of the success-rate in the former than in the latter treatment.

4 Results

4.1 Principals

The data on the principals' estimates confirms Claim 1. Principals in the uninformed treatment are, on average, very well calibrated: there is virtually no difference between their average estimate 39.76% and the true success rate 39.25% ($p = 0.824$).[15] In contrast, principals in the informed treatment grossly overestimate the success rate for the vast majority of tasks.[16] Their average estimate of the success rate amounts to 57.45% which is significantly higher than both the true success rate ($p < 0.001$) and the average estimate of the principals in the uninformed treatment ($p < 0.001$). Accordingly, principals in the informed treatment had significantly lower expected earnings (€ 2.40) than principals in the unformed treatment (€ 3.65; one-sided t-test: $p = 0.034$).[17]

[15] We employed a t-test of the average estimate per principal against the average success rate (over all tasks). Figure 5 in the Appendix shows the distribution of individual performance estimates by informed and uninformed principals together with the actual performance of the reference agents. A Kolmogorov-Smirnov test of the distributions of average individual estimates between treatments yields $p = 0.001$. Unless stated otherwise, p-values throughout the result section refer to (two-sided) t-tests that are based on average values per subject.

[16] Figure 6 in the Appendix provides a plot of the principals' average first-order belief per treatment over time.

[17] The average payoffs in the rounds randomly selected for payment were € 1.50 and € 2.50 in the informed and the uninformed treatment, respectively.

Result 1. *Principals in the informed, but not in the uninformed, treatment overestimate the true success rate on average.*

The above result and the size of the exaggeration is not surprising given the previous findings in the literature, e.g., Loewenstein, Moore and Weber (2006). Reproducing this reinforces the validity of our experiment.[18]

4.2 Agents

4.2.1 Investment decisions

We now move to the agent's investment decision. The data clearly supports Claim 3.[19] Agents matched to informed principals invest at a significantly lower rate than agents matched to uninformed principals.[20] The average investment rate of agents matched to uninformed principals is 67.3%, whereas the average investment rate of agents matched to informed principals is only 39.2% ($p < 0.001$).

Result 2. *Agents invest significantly less often in the informed than in the uninformed treatment.*

The agents in the informed treatment, relative to agents in the uninformed treatment, shy away from choosing an option whose payoff decreases, in the sense of first-order stochastic dominance, in the principal's belief. Result 2 is consistent

[18]Following Moore and Healy (2008), we can also examine the extent to which task difficulty per se plays a role here. If we divide the tasks into hard and easy ones by the median one (yielding 10 hard tasks with success rates of 0.42 and below and 10 easy tasks with success rates of 0.43 and above), we find that principals in the uninformed treatment, on average, overestimate the success rate for hard tasks by 7 percentage points ($p = 0.047$, sign test) but underestimate the true success rate for easy tasks by 6 percentage points ($p = 0.059$). This reversal is well known as the Bayesian hard-easy effect as described by Moore and Healy (2008). This reversal, however, is not observed for principals in the informed treatment. The difference between the informed and uninformed treatments is significant for both easy and hard tasks ($p < 0.01$ for each difficulty level). Principals in the informed treatment significantly overestimate the success rate by 21 percentage points for hard and by 16 percentage points for easy tasks.

[19]Figure 7 in the Appendix shows the distribution of individual investment rates in the informed and the uninformed treatment.

[20]We pool the sessions with belief elicitation and those without. Within the informed [uninformed] treatments, the average investment rates per agent in sessions with belief elicitation do not differ from the average investment rates in sessions without belief elicitation (t-test, $p = 0.76$ [$p = 0.70$]). There are also no significant time trends in the investment rates (see Figure 8 in the Appendix).

with agents anticipating the projection of the principals. Recall that for the agents the only difference between the two treatments is that the agent in the informed treatment was told that his principal had access to the solution while the agent in the uninformed treatment was told that his principal had not been given the solution. Hence, the difference in the propensity to invest has to do with the difference between the agent's first-order and second-order beliefs.[21]

4.3 Stated Beliefs

We now turn to the key structured hypothesis of our study. Figure 3 summarizes our first key findings. It shows a bar chart which collects the average stated estimates of the agents in each treatment together with the true success rate and the corresponding estimates of the principals.

The left panel shows the data from the uninformed treatment. Under projection equilibrium, all beliefs shall be correct on average. This is indeed what we find. None of the elicited estimates are, on average, significantly different from the true success rate: not the agents' first-order estimates ($p = 0.917$), not their second-order estimates ($p = 0.140$), nor the principals' first-order estimates ($p = 0.337$).[22]

The right panel shows the data from the informed treatments. The principals vastly overestimate the true success rate, while the agents' first-order estimates are well calibrated on average ($p = 0.967$), as predicted. The agents' second-order estimates, as predicted, are significantly higher than their *own* first-order estimates ($p < 0.001$) on average. Finally, as predicted, the agents' second-order estimates are also significantly lower than the principals' estimates (one-sided t -test: $p = 0.047$). Agents both anticipate and underestimate the principals' mistaken exaggeration.

When comparing treatments, the agents' first-order estimates are not significantly different ($p = 0.956$) across treatments. The agents' second-order estimate

[21] We find no significant treatment difference in the performance of agents (their success rate is 41.35% in the informed treatment and 39.89% in the uninformed treatment; $p = 0.573$). Thus, any treatment differences in the agents' investment decision or the agents' beliefs cannot be attributed to differences in task performance.

[22] In the informed treatment, the agents' second-order estimates are somewhat higher than the principals' first-order estimates, but this difference is not significant either ($p = 0.080$).

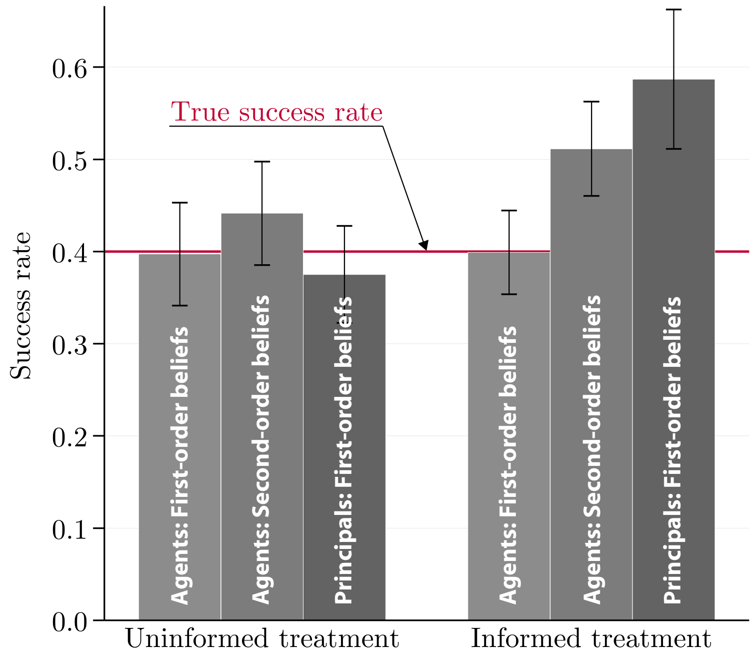


Figure 3: Agents' first-order estimates (estimates of the success rate) and second-order estimates (estimates of the principals' estimate), conditional on being matched with informed or uninformed principals. Capped spikes represent 95% confidence intervals.

in the informed treatment is significantly higher than the same in the uninformed treatment ($p = 0.0314$ for one-sided t-test). Finally, the difference between the agents' second- and own first-order estimates is also significantly larger in the informed than in the uninformed treatment ($p < 0.001$).[23] In sum, the following results hold for the agent's beliefs.

[23]The size of the treatment effect on the principals' estimates and also on the wedge between the agents' first- and second-order estimates is unchanged when controlling for successful task performance. The treatment difference is also robust to controlling for individual characteristics (see Tables 5 and 6 in the Appendix, respectively).

Result 3 (Partial anticipation of information projection)**.** *The following results hold:*

1. *The agent's first-order estimate about the success rate is correct on average in both treatments.*
2. *The difference between the agent's second-order estimate and her own first-order estimate is significantly larger in the informed than in the uninformed treatment.*
3. *In the informed treatment, the agent's estimate of the principal's estimate is higher than the agent's own estimate and lower than the principal's estimate.*

The evidence clearly violates the predictions of the unbiased beliefs supporting an unbiased BNE, but confirms all predictions on the structure of biased beliefs postulated under projection equilibrium.

4.4 Estimation of Projection Equilibrium

The analysis has confirmed all four predictions on the structure of biased beliefs under projection equilibrium. We now turn to the even tighter structure implied by homogeneous projection. Recall that the key aspect of the parsimony of the model is that the extent of first-degree projection fully determines all higher-order implications and in particular the extent of her awareness of the biases of others, e.g., a player's second-degree projection. The former is recoverable from a player's misprediction of differentially-informed others' first-order beliefs, the latter from a player's misprediction of differentially-informed others' prediction of her first-order beliefs. In our design, the former is inferred from the choices of the principals and the latter from the choices of the agents. Under heterogeneous projection, i.e., $\rho_A \neq \rho_P$, these two can then differ substantially, and the prediction is simply that the agent underestimates the principal's exaggeration. Under homogeneous projection, however, i.e., $\rho = \rho_A = \rho_P$, these two must match and the extent to the principal's exaggeration needs to fully match the extent of the agent's underestimation of this exaggeration.

To test this key aspect of the model, we first use the aggregate data to solve Claim 1 and 2 under heterogeneous projection. This yields the unique solution of $\widehat{\rho}_P = 0.3$ and $\widehat{\rho}_A = 0.33$. That is, the degree of projection inferred from the mistake in the principals' second-order beliefs and the degree of projection inferred from the mistake in the agents' third-order beliefs are remarkably close to being homogeneous. The extent to which the principal exxaggerates the success rate is very close to being the square-root of the extent to which the agent underestimates this exaggeration.

4.4.1 An Econometric Analysis

We now turn to an econometric test of the hypothesis of homogenous projection. We employ a Maximum Likelihood estimation of a random-coefficient model and start with a flexible specification that allows for different degrees of projection both across roles as well as within roles. The parameters of the unrestricted model are $\Theta_{UR} = \{\rho_P, \rho_A, \phi_\rho, \phi_b\}$, where ρ_P and ρ_A denote the average degree of projection in the principal and the agent populations, respectively, and ϕ_ρ and ϕ_b are precision parameters governing variance in individual projection and noise in response, as will become clear in a moment. We then estimate the model under homogeneous projection, i.e., with restricted parameters $\Theta_R = \{\rho_P = \rho_A, \phi_\rho, \phi_b\}$. In our design, a comparison of the restricted and the unrestricted specification provides the ultimate test since it directly ties together the extent of the basic mistake with the extent of its anticipation in others.

Our econometric model makes repeated use of the beta distribution. We will use a convenient alternative parameterization of the beta distribution $x \sim \text{Beta}(\mu, \phi)$ with density

$$f(x; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu)\Gamma(\phi(1-\mu))} x^{\phi\mu - 1}(1-x)^{\phi(1-\mu)-1},$$

where the first parameter μ is the expected value of x and the second parameter ϕ is a precision parameter that is inversely related to the variance of x, $\text{var}(x) = \mu(1-\mu)/(1+\phi)$ (Ferrari and Cribari-Neto, 2004). That is, conditional on the mean μ, higher values of ϕ translate into a lower variance.

Our first structural assumption concerns the errors subjects make in their estimates. In the following we assume that subjects stated estimates are beta distributed centered around the task- and individual-specific expectations given by Claim 1 and Claim 2. Specifically, the estimate of principal i and agent j in period t are

$$\begin{aligned} b^{I}_{P_i t} &\sim \mathrm{Beta}(\mu_{P_i t}, \phi_b), \\ b^{II}_{A_j t} &\sim \mathrm{Beta}(\mu_{A_j t}, \phi_b), \end{aligned} \tag{SA1}$$

where

$$\mu_{P_i t} = \rho_{P_i} + (1-\rho_{P_i})\pi_t, \qquad \text{(see Claim 1)}$$

$$\mu_{A_j t} = \pi_t + (1-\rho_{A_j})\rho_P(1-\pi_t). \qquad \text{(see Claim 2)}$$

Our second structural assumption captures unobserved individual heterogeneity in the degree of informational projection and accounts for repeated observations on the individual level. We employ a random-coefficient specification where individual degrees projection in the principal and the agent populations follow beta distributions with

$$\begin{aligned} \rho_{P_i} &\sim \mathrm{Beta}(\rho_P, \phi_\rho), \\ \rho_{A_j} &\sim \mathrm{Beta}(\rho_A, \phi_\rho). \end{aligned} \tag{SA2}$$

We impose tight restrictions on the distributions of the degree of projection in the agent and the principal populations by allowing them to differ only with respect to their location parameter. This greatly facilitates our test of equality of the average degree of projection across roles which is in the focus of this section. We take a closer look at heterogeneity and provide a test of this structural assumption in section 4.5.

We now formulate the log-likelihood function. Conditional on ρ_{k_i} and ϕ_ρ, the likelihood of observing the sequence of stated estimates $(b_{k_i t})_t$ of subject i in role $k \in \{A, P\}$ is given by

$$L_{k_i}(\rho_{k_i}, \phi_b) = \textstyle\prod_t f_b\left(b_{k_i t}; \mu_{k_i t}(\rho_{k_i}), \phi_b\right).$$

Hence, the unconditional probability amounts to

$$L_{k_i}(\rho_k, \phi_\rho, \phi_b) = \int \left[\prod_t f_b\left(b_{k_it}; \mu_{k_it}(\rho_{k_i}), \phi_b\right)\right] f_\rho\left(\rho_{k_i}; \rho_k, \phi_\rho\right) d\rho_{k_i}. \tag{1}$$

The joint log likelihood function of the principals' and the agents' responses can then be written as

$$l(\rho_P, \rho_A, \phi_\rho, \phi_b) = \sum_k \sum_i \log L_{k_i}(\rho_k, \phi_\rho, \phi_b). \tag{2}$$

We estimate the parameters in (2) by maximum simulated likelihood (Train 2009; Wooldridge, 2010).[24] Table 1 shows the estimation results for the unrestricted model with $\rho_P \neq \rho_A$ in the left column and the restricted model with $\rho_P = \rho_A$ in the right column.[25] We focus on the unrestricted model first where we make three observations.

First, the principals' average degree of projection is estimated to be $\widehat{\rho}_P = 0.326$ with a confidence interval of $[0.247, 0.405]$. This estimate clearly indicates the relevance of informational projection: the unbiased BNE—which is the special case where ρ_P is zero—is clearly rejected. Second, the agents' average degree of projection, the extent to which the agent under-appreciates the principal's bias is estimated to be 0.334 with a confidence interval of $[0.110, 0.558]$. The $\widehat{\rho}_A = 0.334$ estimate—which is significantly different from 0 and 1—gives structure to our observation that agents do anticipate that the principals are partially biased—but under-anticipate the principals' level of projection.

Crucially, the estimated parameters of the degree of projection are not significantly different between the principals and the agents ($p = 0.914$). Furthermore, the log likelihood of the unrestricted model is very close to the one of the restricted model (right column of Table 1), and standard model selection criteria (e.g., BIC) clearly favor the single-parameter model of homogeneous projection over the un-

[24]The estimation is conducted with GAUSS. We use Halton sequences of length $R = 100,000$ for each individual with different primes as the basis for the sequences for the principals and the agents (see Train 2009, p221ff).

[25]The results are robust with respect to alternative starting values for the estimation procedure. All regressions for a uniform grid of starting values converge to the same estimates (both for the restricted and the unrestricted model). Thus, the likelihood function in (2) appears to assume a global (and unique) maximum at the estimated parameters.

Table 1: Maximum likelihood estimates of projection bias ρ based on Claim 1 and 2.

	Unrestricted model with heterogeneous ρ ($\rho_P \neq \rho_A$)		Restricted model with homogeneous ρ ($\rho_P = \rho_A$)	
Parameter	Estimate	Conf. interval	Estimate	Conf. interval
ρ_P	0.326^{***}	$[0.247, 0.405]$	0.324^{***}	$[0.252, 0.397]$
ρ_A	0.334^{**}	$[0.110, 0.558]$		
ϕ_ρ	3.103^{***}	$[1.169, 5.036]$	3.092^{***}	$[1.177, 5.007]$
ϕ_b	4.377^{***}	$[3.938, 4.815]$	4.377^{***}	$[3.939, 4.815]$
N	720		720	
$\ln L$	123.597		123.593	

Note: Values in square brackets represent 95% confidence intervals. Asterisks represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ Testing $H_0 : \rho_P = \rho_A$ in column (1) yields $p = 0.8389$.

restricted model with two parameters. In short, the data is remarkably consistent with the structure of biased beliefs implied by projection equilibrium, that is, the tight link between the basic mistake and the mistake in the anticipation of this basic mistake in others.

4.5 Partial Bias and Partial Anticipation

The final, but crucial, part of the analysis is devoted to exploring individual degrees of informational projection. This analysis provides the ultimate test of the idea of a partial bias and the underlying notion of a coherent but mistaken egocentric belief hierarchy with partial adjustment to the truth at the *individual level.* This is done in conjunction with a test of the econometric specification with regard to the structural assumption (SA2) made above. Specifically, we test whether the mean degree of projection $\widehat{\rho} = 0.324$ estimated from (2) is indeed generated by a beta distribution of ρ_{k_i}. A misspecification in this matter would not only be relevant from an econometric point of view; it may also challenge the interpretation of our results. In particular, if the estimated average degree of projection is the result of a mixture of some people being naive and not anticipating the mistakes of others at all ($\rho = 1$) and others being sophisticated fully anticipating it ($\rho = 0$), then Claim 2 and the logic of partial anticipation and partial adjustment would have no descriptive accuracy on the individual level.

We base our specification test on non-parametric density estimates of individual degrees of projection in the principal and agent populations. To this end, we first obtain individual estimates of ρ for each principal and each agent from the informed treatment using simple linear regressions without imposing any restrictions on the size or the sign of the parameters. In contrast to the previous subsection, we now adopt a simple OLS framework and use the subjects' conditional estimates, that is, their estimates conditional on their own performance in the task. Specifically, for each principal i in the informed treatment, we estimate her degree of projection ρ_{P_i} from Claim 1 via:

$$b^{I}_{P_i t} = \pi_t + \rho_{P_i}(1 - \pi_t) + \epsilon_{it}, \tag{3}$$

where ϵ_{it} denotes an error term with mean zero and variance σ_i^2. Analogously, for each agent j in the informed treatment we estimate his degree of projection ρ_{A_j} from Claim 2 via

$$b^{II}_{A_j t} = b^{I}_{A_j t} + (1 - \rho_{A_j})\rho_P(1 - \pi_t) + \epsilon_{jt}, \tag{4}$$

where ρ_P is the mean projection bias in the principal population, and ϵ_{jt} denotes an error term with mean zero and variance σ_j^2. For a derivation of these equations see the Appendix.[26] We estimate the parameters in (3) and (4) by OLS, where we substitute ρ_P in (4) with the average estimate of ρ_{P_i} obtained from the regressions in (3).[27]

[26]The results are very similar and qualitatively the same when using the agents' first-order estimates instead of the true success rates in (4). Figure 13 in the appendix shows the corresponding distribution of individual ρ estimates for a comparison with Figure 4.

[27]We base all inference on the individual level on heteroskedasticity-robust standard errors. Unlike in the simultaneous estimation of the agents' and the principals' projection bias from (2), the simple estimation approach applied here assures that the individual estimates of the principals' projection bias are not informed by the data of the agents' choices, a feature that is desirable for our specification test below. Figure 11 in the Appendix plots the average predicted second-order estimate together with the actual second-order estimates of the agents.

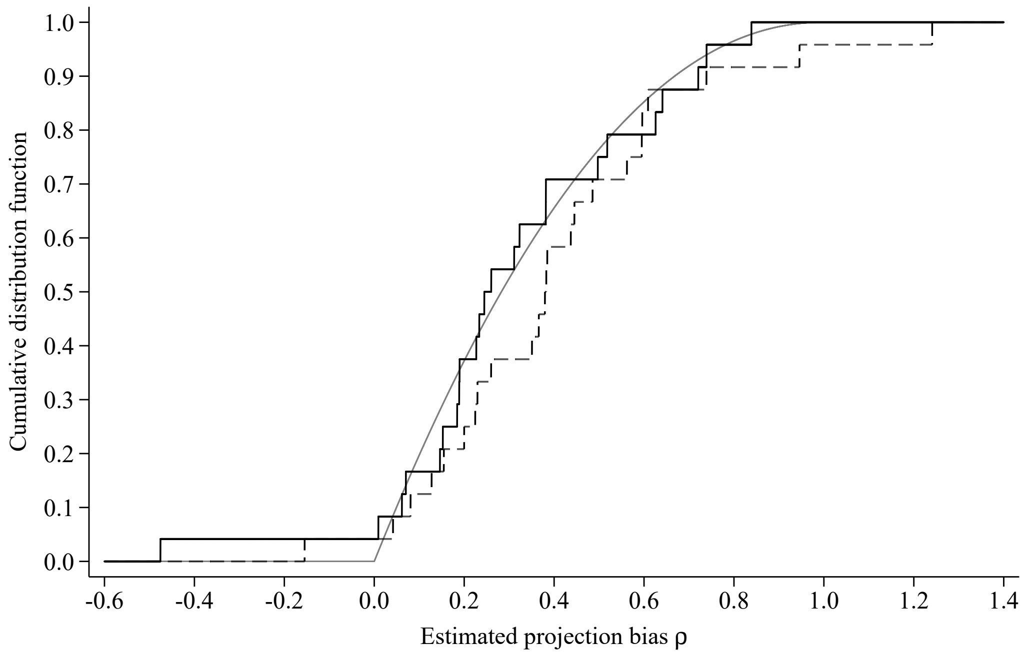


Figure 4: Empirical cumulative distribution functions of principals' (solid) and agents' (dashed) projection bias ρ in the informed treatment. The smooth line shows the estimated beta distribution from the model with homogeneous projection (right column in Table 1).

Figure 4 shows the empirical CDFs of the individual degrees of projection in the principal and the agent populations. A casual inspection of the figure already suggests that the empirical CDFs of the principals' and the agents' ρs are quite similar. In fact, a Kolmogorov-Smirnov test does not reveal any significant difference between the distributions ($p = 0.441$). That is, not only are the average degrees of projection exhibited by the principals, inferred from their basic mistake, and the agents, inferred from the extent of their under-appreciation of the principals' basic mistake, are the same, but the two distributions describing the underlying heterogeneity and the distribution of partialness in one and the other are also not significantly different.

Second, and crucially, as is visible on Figure 4, partial projection is the norm both when it comes to first-degree projection and second-degree projection (the mistake in the anticipation). The majority of the principals (**70.8**%) exhibit an estimated ρ that is significantly larger than zero and significantly smaller than one. Similarly, the majority of the agents (**50**%) have an estimated ρ_{A_j} that is significantly larger than zero, but significantly smaller than one.[28] People do believe that others are partially biased and partially underestimate its extent.

Finally, Figure 4 conveys an additional result regarding the econometric specification in (2). The pooled empirical CDF of the principals' and the agents' projection is not significantly different from the beta distribution $f_\rho(\widehat{\rho} = 0.324, \widehat{\phi}_\rho = 3.092)$ obtained from model (2) with homogeneous projection (smooth line in Figure 4; Kolmogorov-Smirnov test: $p = 0.529$).[29]

In sum, the findings lend surprisingly strong support to the most parsimonious specification of projection equilibrium and, perhaps more importantly, to the realism of underlying logic a partial belief bias.

5 Discussion

We are unaware of any other existing model of strategic behavior that would provide a tight explanation of the data. Below we describe the implications of some leading candidates from the literature.

5.1 Alternative Models and Mechanisms

Coarse Thinking. Unlike a number of other prominent behavioral models of play in games with private information, projection equilibrium focuses on players misperceiving others' beliefs, and having wrong explicit beliefs about the beliefs

[28] The second most common category (33.3%) in the agent population is ρ being not significantly different from 0 but significantly different from 1, i.e., full anticipation of others' information projection. Further 12.5% of the agents have an estimated ρ that is not significantly different from 1 but significantly different from 0, i.e., no anticipation of the principals' information projection. One agent (the remaining 4.2% of the agents) has an estimated ρ that is significantly larger than one. The corresponding fractions for the principals are similar and there is no significant difference between the agents' and the principals' categorized distribution of projection bias (Fisher's exact test: $p = 0.179$).

[29] The separate empirical CDFs principals' and the agents' degree of projection are also well described by beta distributions Kolmogorov-Smirnov tests of the empirical CDF against the best-fitting beta distribution yields $p = 0.941$ for the principals and $p = 0.974$ for the agents.

of others rather than misperceiving the relationship between other players' beliefs and their actions. In particular, the models of ABEE (Jehiel, 2005), and cursed equilibrium (Eyster and Rabin, 2005), assume that people have correct expectations about the information of others, but have coarse or misspecified expectations about the link between others' actions and their information. Crucially, these models are then closed by the identifying assumption that those expectations are nevertheless correct on average, that is, each player has correct expectations about the distribution of her opponent's actions.

The above identifying assumption directly implies that in our design, both models have the same overall predictions as the unbiased BNE. In the context of the current experiment, they both imply a null treatment effect. A principal should never exaggerate the agent's performance on average and the agent should never anticipate any mistake by the principal on average.[30]

Note also, that in these alternative models, unlike under projection equilibrium, people need not have coherent beliefs about the beliefs of others. For example, for a partially cursed player it may not be possible to justify her expectation about her opponent's behavior based on some belief about his beliefs and his rationality, e.g., she has to be believe that he acts irrationally given his information and with some probability.

Risk Aversion. We find no evidence that risk aversion matters for the subjects' choices. (See Tables 3 and 6 in the Appendix). Note also that if more information helps unbiased principals to make more accurate forecasts on average, under unbiased beliefs, or the absence of any anticipation of the projection of others, and risk aversion, the agent should be choosing the risky option over the safe option more often when the principal is informed rather than when she is not. Instead we find the exact opposite pattern.

Overconfidence. Note that overconfidence cannot explain the subjects' choices either. If an agent believes that she is better than average, then she might underestimate the reference agents' performance relative to her own, but this will not differ across treatments. Furthermore, as the data shows, both of

[30]Note also that QRE also predicts no treatment difference since the principal's incentives in the two treatments are exactly the same. The same is true for level-k models that hold the level zero play constant across treatments.

these estimates, and also that of the principals in the uninformed treatment, are in fact unbiased. Similarly, a principal may be over- or under-confident when inferring about others' performance on a given task, but there is no reason for this to systematically interact with the treatment.

Everybody is just like me. Finally, one may propose a general heuristic whereby people simply think that others are just like them, whatever this may mean. While the exact meaning of such a heuristic may be unclear, note that if people just believed that others have the same beliefs as they do, then we cannot account for our key finding; the systematic wedge between the majority of the agents' own first-order and second-order beliefs in the informed treatment, that is, the fact that the typical subject explicitly thinks that others form systematically wrong beliefs about her beliefs. The same then holds a fortiori about the data providing support for partial bias at the individual level.

5.2 Conditional Estimates

Note that in the absence of further assumptions we cannot pin down the wedge between the unbiased and biased conditional estimates, that is players' estimates conditional on whether or not they themselves were able to solve the task because we do not know the unbiased conditional estimates only the unbiased average estimate. At the same time, players' conditional beliefs within a treatment, that is, conditional on whether or not they themselves were able to solve the task, shall also be affected by their bias under projection equilibrium. In particular, within the uninformed treatment, a principal who figures out the solution herself, by projecting her information, shall comparatively exaggerate the success rate of others and a player who does not figure it out, by projecting her ignorance, shall comparatively underestimate the success rate. Projection thus inflates the difference between these conditional estimates. Consistent with this, in the uninformed treatment the principal's estimate of the success rate conditional on spotting the difference was 60.93% while the same conditional on the principal not figuring out the solution was 29.95%.[31]

[31] Similarly, the agents' average estimate of the success rate is 53.41% when they found the solution but only 32.86%when they did not.

More importantly, we can compare the estimates of the principals in the informed treatment who were given the solution to the estimates of the principals in the uninformed treatment who were not given the solution but who did figure this out themselves. If any systematic distortion in conditional estimates is due to informational projection only, then these two should be the same. We do find that the estimates of the principals who spotted the difference in the uninformed treatment (60.93%) is very close to the estimates of the principals who were given the solution in the informed treatment (57.45%). This finding provides further support for our basic premise that the distortion in the principals' estimates is due to informational projection as opposed to some alternative psychological mechanism whose implications would greatly differ in the way information is acquired, e.g., problems that one solves may appear more difficult while problems for which one is exogenously given the solution just appear too easy.

5.3 Conclusion

This paper studies people's perception of the biases of others. While a host of robust findings demonstrate that people engage in limited informational perspective taking the very meaning and implications of such a social bias in strategic settings crucially depend on the extent to which people simultaneously also anticipate this tendency in each other. Our study lends surprisingly strong empirical support to the parsimonious model of projection equilibrium which, by proposing a mode of partial bias, postulates a tight link between these two. It also confirms the realism of the underlying notion of a partial belief bias. Most people are neither fully naive nor sophisticated about the biases of others. Instead they explicitly believe that others are partially biased but, in proportion to their own bias, underestimate its extent in others.

Our results also illustrates well the potential of obtaining key novel insights by eliciting higher-order beliefs in studying models of biased social cognition. While there are multiple factors that might impact strategic behavior and help capture departures from classic rational choice outcomes in social settings, it is the pattern in higher-order beliefs, beliefs about those biases, that may be essential in providing a more careful account. For example, if players had biased beliefs about others (informed players exaggerated the performance of uninformed players) because everyone thinks others are just like them, then it is impossible to account for the partial anticipation of this exaggeration by the agents, as would be manifest in defensive medicine or successful deception, as predicted by projection equilibrium even in homogenous populations, and established by our findings. More generally, eliciting higher-order beliefs help better understand the nature of biased cognition and help guide modeling choices. For example, in the case of cursed equilibrium, Eyster and Rabin (2005) players may not be able to hold coherent beliefs about how others, e,g., may hold a mixture of incoherent and correct beliefs. Instead we find that a fully coherent but parsimoniously and significantly biased belief hierarchy provides a very good description of the actual data. The findings may also help shed light on the robustness of the basic mistake to debiasing method, e.g., Wu et al. (2010). People may well be aware of a mistaken tendency such as hindsight bias and learn to anticipate it in others, yet consistently continue to suffer from it themselves.

Providing empirical support for projection equilibrium, which includes the idea that measuring the basic mistake may be sufficient to pin down the entire structure of the bias and the simultaneous presence of the basic mistake and limited learning about the mistakes of others, is potentially helpful for understanding various economic problems such as the relationship between information and incentives (e.g., Holmström 1979), the link between risk taking and the allocation of responsibility or liability in agency settings (e.g., Harley 2007 argues that hindsight bias is a key factor in the judgement of jurors in courts), or the functioning of authority in organizations (Aghion and Tirole, 1997). In all these settings the basic mistake and its anticipation in others jointly matter. Our findings are also consistent with the application of projection equilibrium to other strategic settings. In particular, in the classic context of bilateral trade with common values and private information, Madarasz (2016) finds that the model provides a very close fit of the experimental data (e.g., Samuelson and Bazerman, 1985; Holt and Sherman, 1994).[32] Future research can explore the extent to which the anticipation of the biases of others

[32]The degree of the bias which explains the aggregate empirical findings in that context is consistent with the extent of information projection found in the current study.

is directly present also in other social contexts (e.g., Mobius et al. 2014) and the relationship between such anticipation and a basic mistake itself.

References

[1] Aghion, P. and J. Tirole (1997): "Formal and Real Authority in Organizations." *Journal of Political Economy*, 105(1): 1-29.

[2] Augenblick, N and M. Rabin (2018): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *Review of Economic Studies*, forthcoming.

[3] Baron, J., and J. Hershey (1988): "Outcome Bias in Decision Evaluation." *Journal of Personality and Social Psychology*, 54(4), 569–579.

[4] Benabou, R., and J. Tirole (2016): "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, 30(3), 141–164.

[5] Berlin L. (2003): Statement of Leonard Berlin, M.D., to the U.S. Senate Committee on Health, Education Labor and Pensions: Mammography Quality Standards Act Reauthorization. http://www.fda.gov/ohrms/dockets/ac/03/briefing/3945b1_05_Berlin%20testimony.pdf.

[6] Bhatt, M., and C. F. Camerer. (2005): "Self-referential Thinking and Equilibrium as States of Mind in Games: fMRI Evidence." *Games and Economic Behavior*, 52(2), 424–459.

[7] Blanco, M., Engelmann, D., Koch, A.K., and H.-T. Norman. (2010): "Belief Elicitation in Experiments: Is There a Hedging Problem?" *Experimental Economics*, 13(4), 412–438.

[8] Buehler, R., Griffin, D., and Ross, M. (1994): "Exploring the "planning fallacy": Why people underestimate their task completion times." *Journal of Personality and Social Psychology*, 67(3), 366–381.

[9] Camerer, C., Loewenstein, G., and M. Weber. (1989): "The Curse of Knowledge in Economic Settings: An Experimental Analysis." *Journal of Political Economy*, 97(5), 1232–1254.

[10] Cason, T., and C. Plott. (2014): "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy*, 122(6), 1235–1270.

[11] Danz, D. (2014): "The Curse of Knowledge Increases Self-Selection into Competition: Experimental Evidence." *Working paper SPII 2014-207*, WZB Berlin Social Science Center.

[12] Davis, M. H. (1983): "Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach." *Journal of Personality and Social Psychology*, 44(1), 113–126.

[13] Eyster, E., and M. Rabin. (2005): "Cursed equilibrium." *Econometrica*, 73(5), 1623–1672.

[14] Ferrari, S., and F. Cribari-Neto. (2004): "Beta regression for modelling rates and proportions." *Journal of Applied Statistics*, 31(7), 799–815.

[15] Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics*, 10(2), 171–178.

[16] Fischhoff, B. (1975): "Hindsight / foresight: The Effect of Outcome Knowledge On Judgement Under Uncertainty." *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288–299.

[17] Gennaioli, N., and A. Shleifer. (2010): "What Comes to Mind? *Quarterly Journal of Economics*, 125(4), 1399–1433.

[18] Gilovich, T., Savitsky, K., and V. Medvec. (1998): "The Illusion of Transparency: Biased Assessment of Other's Ability to Read our Emotional States." *Journal of Personality and Social Psychology*, 75(2), 332–346.

[19] Greiner, B. (2004): "An Online Recruitment System for Economic Experiments." in *Forschung und wissenschaftliches Rechnen 2003*, ed. by K. Kremer

and V. Macho. GWDG Bericht 63, Göttingen: Ges. für Wiss. Datenverarbeitung.

[20] Holmström, B. (1979): "Moral Hazard and Observability." *The Bell Journal of Economics*, 10(1), 74–91.

[21] Holt, C., and Sherman, R. (1994): "The loser's curse,"*American Economic Review*, 84(3), 642–652.

[22] Jehiel, P., and F. Koessler (2008): "Revisiting games of incomplete information with analogy-based expectations." *Games and Economic Behavior*, 62(2), 533–557.

[23] Loewenstein, G., Moore, D., and R. Weber. (2006): "Misperceiving the Value of Information in Predicting the Performance of Others." *Experimental Economics*, 9(3), 281–95.

[24] Madarász, K. (2012): "Information Projection: Model and Applications." *Review of Economic Studies*, 79(3), 961–985.

[25] Madarász, K. (2014): "Projection Equilibrium: Definition and Applications to Social Investment, Communication and Trade." CEPR D.P, revised 2016, https://works.bepress.com/kristof_madarasz/43/.

[26] Manski, C., and C. Neri. (2013): "First- and Second-order Subjective Expectations in Strategic Decision-making: Experimental Evidence." *Games and Economic Behavior*, 81, 232–254.

[27] Möbius, M., M. Niederle, P. Niehaus, and T. S. Rosenblat. (2014): "Managing Self-Confidence." *Working Paper.*

[28] Moore, D. and Healy, P.J. (2008): "The trouble with overconfidence." *Psychological Review*, 115(2), 502–517.

[29] Morris, S. and H. S. Shin (2006). "Global Games: Theory and Applications." *In Advances in Economics and Econometrics, Eds. M. Dewatripont, L. P. Hansen, S.Turnovsky.* Cambridge University Press

[30] Piaget, J. 1952: "The Origins of Intelligence in Children." International Universities Press, New York.

[31] Rammstedt, B., and O. P. John. (2007): "Measuring Personality in One Minute or Less: A 10-item Short Version of the Big Five Inventory in English and German." *Journal of Research in Personality*, 41(1), 203–212.

[32] Rensink, R. A., O'Regan, J. K., and J. J. Clark .(1997): "To See or Not to See: The Need for Attention to Perceive Changes in Scenes." *Psychological Science*, 8(5), 368–373.

[33] Samuelson, W.F., and Bazerman, M.H. (1985): "Negotiation under the winner's curse,"*Research in experimental economics*, 3, 105–38.

[34] Schotter, A., and I. Trevino .(2014): "Belief Elicitation in the Laboratory." *Annual Review of Economics*, 6, 103–128.

[35] Simons, D. J., and D. T. Levin. (1997): "Change Blindness." *Trends in Cognitive Sciences*, 1(7), 261–267.

[36] Thaler, R. (2015): "Misbehaving: The Making of Behavioral Economics." W.W. Norton.

[37] Train, K. (2009): "Discrete choice methods with simulation." Cambridge University Press.

[38] Tversky, A., and D. Kahneman. (1974): "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157), 1124–1131.

[39] Van Boven, L., Loewenstein, G. and Dunning, D. (2003): "Mispredicting the endowment effect: Underestimation of owners' selling prices by buyer's agents." *Journal of Economic Behavior & Organization*, 51(3), 351–365.

[40] Vuong, Q. H. (1989): "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica*, 57(2), 307—333.

[41] Wang, S. W., Filiba, M., and C. Camerer (2010): "Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters." *Working Paper.*

[42] Wimmer H., and Perner J. (1983): "Beliefs about beliefs: Representation and constraircing function of wrong beliefs in young children's understanding of deception." Cognition: 102–128.

[43] Wooldridge, J. M. (2010): "Econometric analysis of cross section and panel data," MIT Press.

[44] Wu, D.-A., Shimojo, S., Wang, S. W., and C. Camerer. (2012): "Shared Visual Attention Reduces Hindsight Bias." *Psychological Science*, 23(12), 1524–1533.

Appendix

5.4 Proofs for Section 3

We now formally present the predictions based on projection equilibrium stated in Madarasz (2016).[33] Since the reference agents solving the task have no relevant strategic interactions, we can introduce a representative reference agent and denote it by $\overline{A}$. This is a short-hand for the ex ante expected average performance of the population of reference agents. With a slight abuse of notation, we can then represent the average success rate of the reference agents in a realized state ω by $\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)]]$. The ex-ante expectation of this is then $E_\omega[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)]] = \pi$. Finally, throughout the analysis we assume that all estimates by all players are formed at the time of solving the basic task, that is, prior to any feed-back. Below, E refers to the expectations operator with respect to the true distribution of actions and signals in the game.

Let E^{ρ_P} denote the expectations of a ρ_P-biased principal. The ex-ante expected estimate of π by the principal is:

$$E_\omega\left[E^{\rho_P}[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)]\right].$$

Using the definition of projection equilibrium, we obtain that:

$$E_\omega\left[\rho_P \max_{x\in D} E[f(\omega,x) \mid P_P(\omega)] + (1-\rho_P)E[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)]\right]$$

This then becomes:

$$\rho_P E_\omega\left[\max_{x\in D} E[f(\omega,x) \mid P_P(\omega)]\right] + (1-\rho_P)E[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)],$$

which equals $\rho_P(d+\pi) + (1-\rho_P)\pi = \pi + \rho_P d$ as stated in Claim 1.

Let E^{ρ_A} denote the expectations of a ρ_A-biased agent. The ex-ante expected first-order estimate of π by the agent is:

[33] See Section 5 of Madarasz (2016).

$$
\begin{aligned}
& E_\omega\left[E^{\rho_A}[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_A(\omega)]\right] \\
= \quad & \rho_A E_\omega\left[\max_{x\in D} E[f(\omega,x) \mid P_A(\omega)]\right] + (1-\rho_A)E[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_A(\omega)].
\end{aligned}
$$

This then becomes $\rho_A\pi + (1-\rho_A)\pi = \pi$ which establishes the first part of Claim 2.

Consider now the agent's ex-ante expected second-order estimate, the estimate of the principal's estimate of π,

$$
E_\omega\left[E^{\rho_A}[E^{\rho_P}[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \mid P_A(\omega)]\right].
$$

Using the definition of projection equilibrium we can re-write this as:

$$
E_\omega\left[\rho_A E[\max_{x\in D} E[f(\omega,x) \mid P_A(\omega)] + (1-\rho_A)E[E^{\rho_P}[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \mid P_A(\omega)]\right]
$$

The first part of the above expression is based on the feature of projection equilibrium that the agent believes that the projected versions of the other players, that is, the projected versions of the reference agents and the principal, all of whom have the same first-order beliefs about the solution to the basic task as the agent does, occur in a perfectly correlated fashion. Furthermore, these respective projected versions of others know that they occur in a perfectly correlated manner. We can then re-arrange the above expression to obtain:

$$
\rho_A\pi + (1-\rho_A)E_\omega[E[E^{\rho_P}[\max_{x\in D} E[f(\omega,x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \mid P_A(\omega)]]
$$

Given Claim 1, the above then equals term equals $\rho_A\pi + (1-\rho_A)(\pi + \rho_P d) = \pi + (1-\rho_A)\rho_P d$ as stated in Claim 2.

5.5 Proofs for Section 4.5

The specification for the principal follows from above. Consider now the agent's estimate conditional on the agent's own success rate. If the agent figures out the solution, her success rate given her own information on the basic task is 1. If the agent does not figure it out, it is some number weakly higher than 0, e.g., random clicking on a 7x7 grid does allow for a positive chance of success. For short, we denote the agent's estimate of her own success rate by the variable [own success]. We can now express the agent's conditional second-order estimate. Under projection equilibrium agent j's stated second-order estimate in task t is then given by:

$$b^{II}_{A_j t} = \rho_{A_j}[\text{own success}] + (1 - \rho_{A_j})E[b^{I}_{P_i t} \mid \text{agent } j\text{'s info}] \tag{5}$$

Note also that

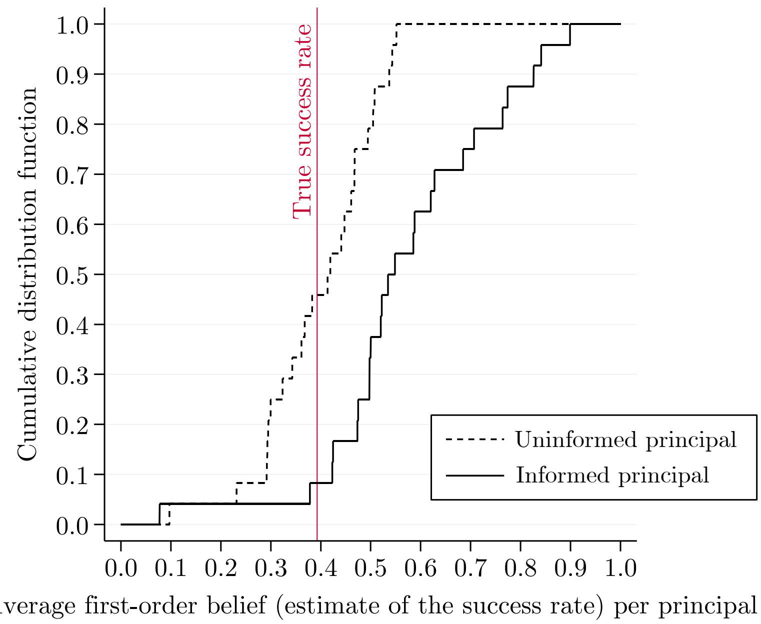
$$b^{I}_{A_j t} = \rho_{A_j}[\text{own success}] + (1 - \rho_{A_j})E[\pi_t \mid \text{agent } j\text{'s info}]. \tag{6}$$

By substituting in Eq(6) into Eq(5), we get that $b^{II}_{A_j t} = b^{I}_{A_j t} - (1 - \rho_{A_j})E[\pi_t \mid \text{agent } j\text{'s info}] + (1 - \rho_{A_j})E[b^{I}_{P_i t} \mid \text{agent } j\text{'s info}]$. Hence, $b^{II}_{A_j t} = b^{I}_{A_j t} - (1 - \rho_{A_j})\pi_t + (1 - \rho_{A_j})(\rho_P + (1 - \rho_P)\pi_t) + \varepsilon_{j,t} = b^{I}_{A_j t} + (1 - \rho_{A_j})\rho_P(1 - \pi_t) + \varepsilon_{j,t}$ where $\varepsilon_{j,t}$ is a mean-zero error term describing the difference between the ex ante expected mean of a random variable and its realization.

5.6 Supplementary analysis

5.6.1 Stated beliefs of the principals

Figure 5: Distribution of average first-order beliefs per principal in the informed and the uninformed treatment.



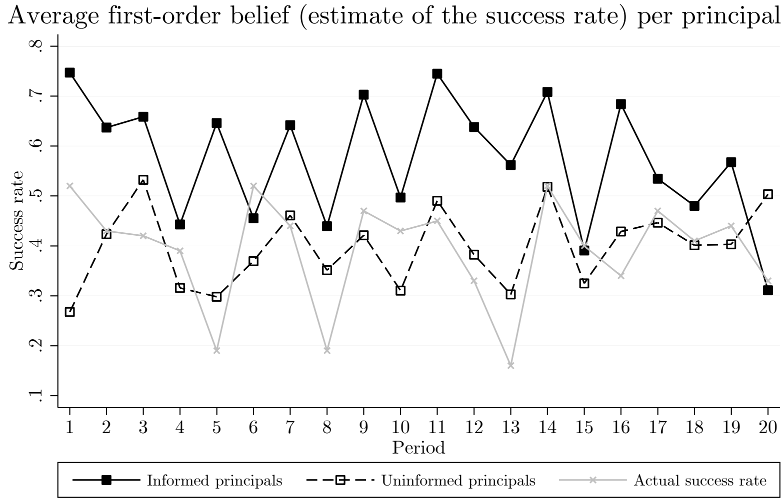
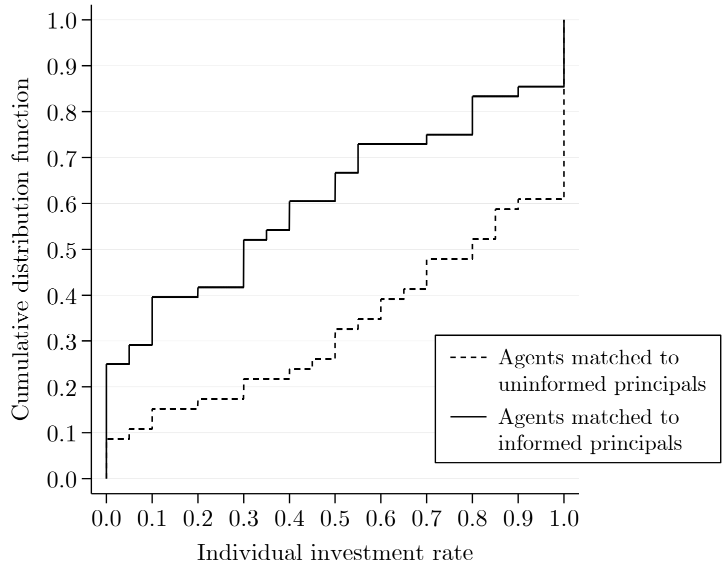


Figure 6: Average performance estimates of principals and actual success rate of the reference agents over time.

5.6.2 Investment decisions of the agents

Figure 7: Distribution of individual investment rates in the informed and the uninformed treatment.



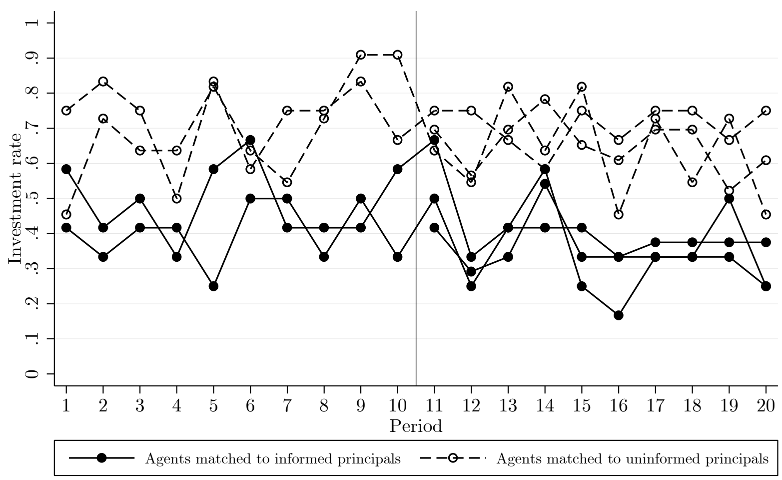


Figure 8: Investment rates per session over time.

Table 2: Propensity to invest conditional on treatment and successful task completion.

Dependent variable (Probit)	Investment decision (1-investment, 0-no investment)		
	(1)	(2)	(3)
Treatment (1-informed)	-0.727^{***} (0.205)	-0.754^{***} (0.211)	-0.738^{***} (0.221)
Success (1-task solved)		0.429^{***} (0.096)	0.451^{***} (0.126)
Treatment×Success			-0.040 (0.190)
Constant	0.467^{***} (0.146)	0.299^{**} (0.149)	0.291^{**} (0.147)
N	1410	1410	1410
R^2	-916.436	-897.552	-897.512
F	12.575	29.023	29.581

Note: Values in parentheses represent standard errors corrected for clusters on the individual level. Asterisks represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3: Regressions of individual investment rates on treatment, gender, and risk attitude.

Dependent variable	Individual investment rate				
(OLS)	(1)	(2)	(3)	(4)	(5)
Treatment (1-informed)	-0.281^{***} (0.075)	-0.279^{***} (0.075)	-0.255^{**} (0.102)	-0.259^{***} (0.077)	-0.254^{**} (0.102)
Gender (1-female)		-0.059 (0.075)	-0.032 (0.108)		-0.048 (0.109)
Treatment×Gender			-0.053 (0.151)		-0.009 (0.157)
Coef. risk aversion (DOSE)				-0.026 (0.022)	-0.024 (0.023)
Constant	0.673^{***} (0.053)	0.698^{***} (0.063)	0.687^{***} (0.071)	0.695^{***} (0.057)	0.715^{***} (0.076)
N	94	94	94	94	94
R^2	0.134	0.140	0.141	0.147	0.151
F	14.230	7.390	4.920	7.813	3.960

Note: Values in parentheses represent standard errors. Asterisks represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5.6.3 Stated beliefs of the agents

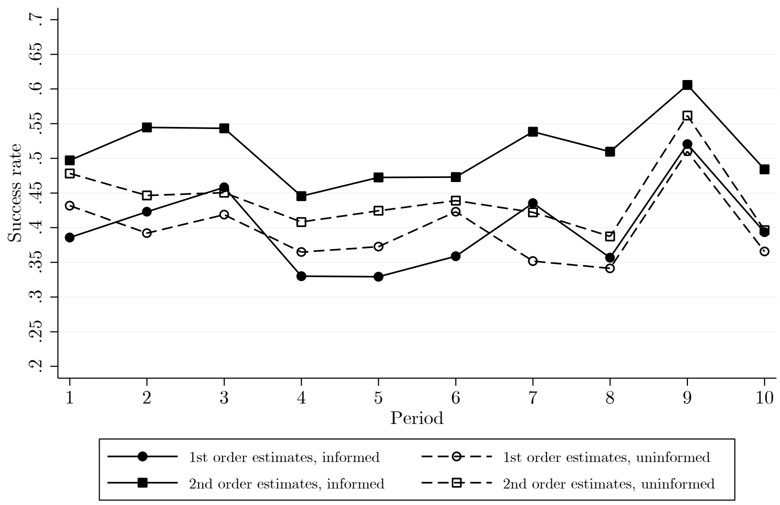


Figure 9: Agents' first-order beliefs (estimates of the success rates of the reference agents) and second-order beliefs (estimates of the principals' estimate) over time, conditional on being matched with informed or uninformed principals.

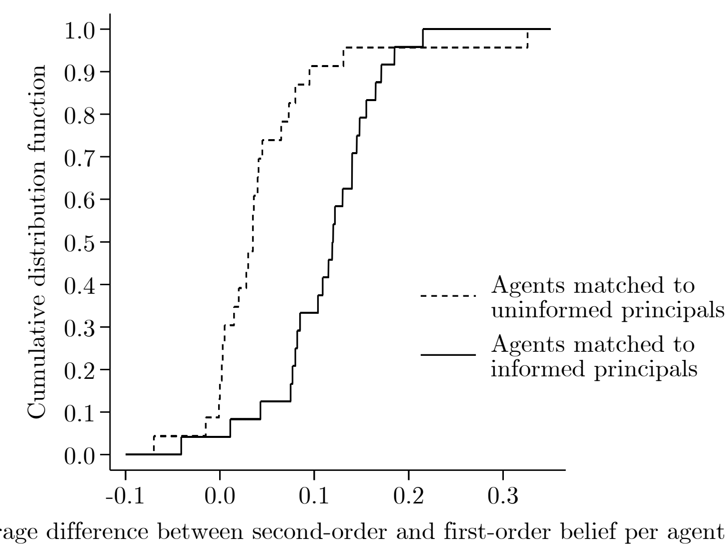


Figure 10: Empirical cumulative distribution functions of each agent's average difference between her second-order belief (estimate of the principal's estimate of the success rate) and her first-order belief (own estimate of the success rates), conditional on being matched with informed or uninformed principals.

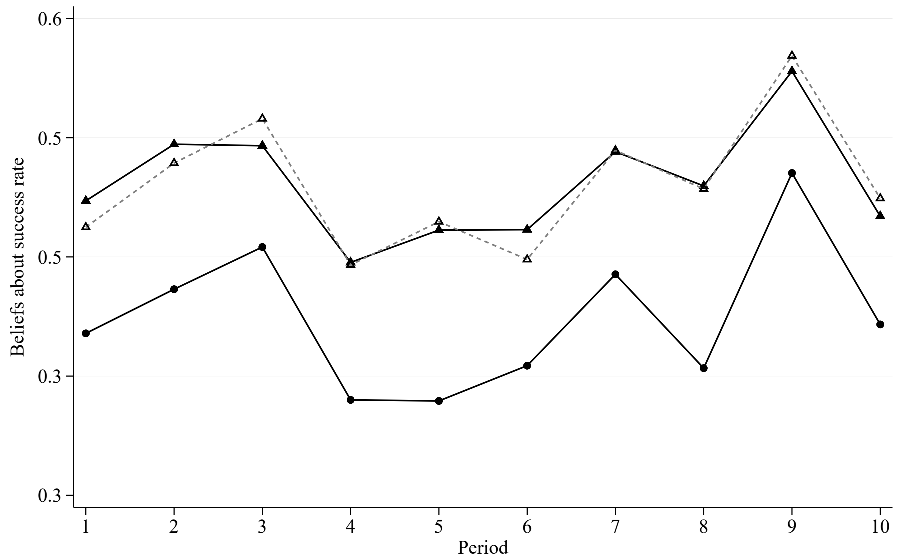


Figure 11: Stated and (average) predicted estimates of the agents. The solid line with circle (triangle) markers shows the agents' average first-order (second-order) estimate per period. The dashed line shows the agents' average second-order estimate per period predicted by individual ρ estimates from equation (4). The line is obtained by predicting each agents' second-order belief for each period and then calculating the average predicted second-order estimate per period.

Table 4: Agents' average first-order beliefs (estimate of the success rates) conditional on treatment and successful task completion.

	(1)	(2)	(3)
Informed	0.002	0.004	0.015
	(0.034)	(0.033)	(0.040)
Success		0.206^{***}	0.222^{***}
		(0.020)	(0.028)
Informed*Success			-0.033
			(0.040)
Constant	0.397^{***}	0.327^{***}	0.321^{***}
	(0.027)	(0.028)	(0.030)
R^2	0.000	0.253	0.255
N	470	470	470

Note: OLS regressions. Values in parentheses are standard errors corrected for clusters on the individual level: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5: Individual differences between second-order beliefs and first-order beliefs conditional on treatment and successful task completion.

Dependent variable	$(b^A_{2,t,i} - b^A_{1,t,i})$		
(OLS)	(1)	(2)	(3)
Treatment	0.068^{***}	0.068^{***}	0.067^{***}
(1-informed)	(0.019)	(0.019)	(0.024)
Success		-0.039^{***}	-0.041^{**}
(1-task solved)		(0.011)	(0.017)
Treatment×Success			0.003
			(0.022)
Constant	0.044^{***}	0.058^{***}	0.058^{***}
	(0.015)	(0.017)	(0.019)
N	470	470	470
R^2	0.090	0.117	0.117
F	12.828	17.767	13.296

Note: Values in parentheses represent standard errors corrected for clusters on the individual level. Asterisks represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6: Mean individual differences in second-order beliefs (estimate of the principal's estimate) and first-order beliefs $b^A_{1,i}$ (estimate of success rate) by treatment and further controls.

Dependent variable	$(b^A_{2,i} - b^A_{1,i}) = T^{-1}\sum_t(b^A_{2,i,t} - b^A_{1,i,t})$				
(OLS)	(1)	(2)	(3)	(4)	(5)
Treatment	0.068^{***}	0.067^{***}	0.089^{***}	0.073^{***}	0.090^{***}
(1-informed)	(0.019)	(0.019)	(0.024)	(0.020)	(0.024)
Gender		0.013	0.047		0.045
(1-female)		(0.020)	(0.029)		(0.030)
Treatment×Gender			−0.062		−0.056
			(0.040)		(0.041)
Coef. risk aversion				−0.006	−0.004
(DOSE)				(0.006)	(0.006)
Constant	0.044^{***}	0.040^{**}	0.030^{*}	0.048^{***}	0.034^{*}
	(0.014)	(0.015)	(0.016)	(0.014)	(0.017)
N	47	47	47	47	47
R^2	0.220	0.228	0.270	0.236	0.278
F	12.720	6.490	5.289	6.787	4.035

Note: Values in parentheses represent standard errors. Asterisk represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

5.7 Estimating projection equiblibrium parameters

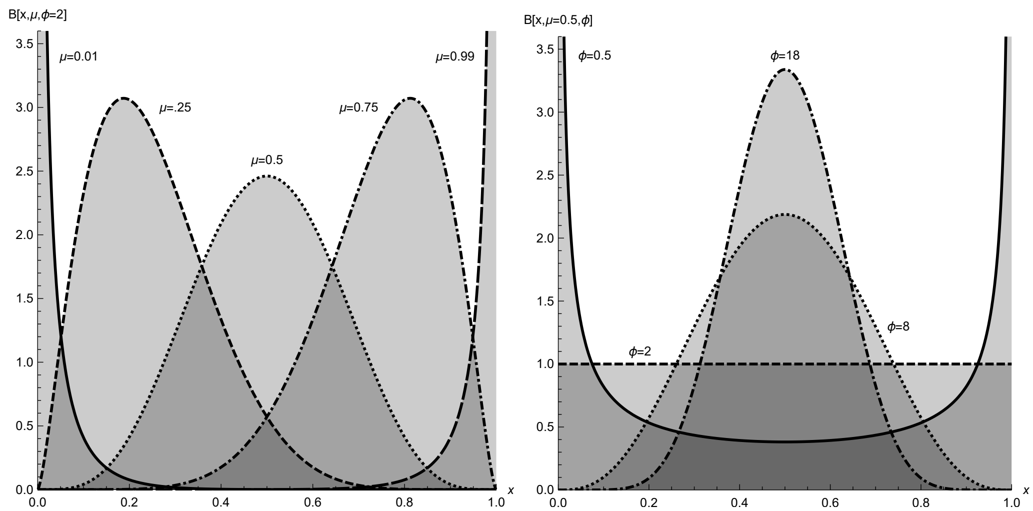


Figure 12: Beta distribution with alternative parameterization $x \sim \text{Beta}(\mu, \phi)$ (Ferrari and Cribari-Neto, 2004).

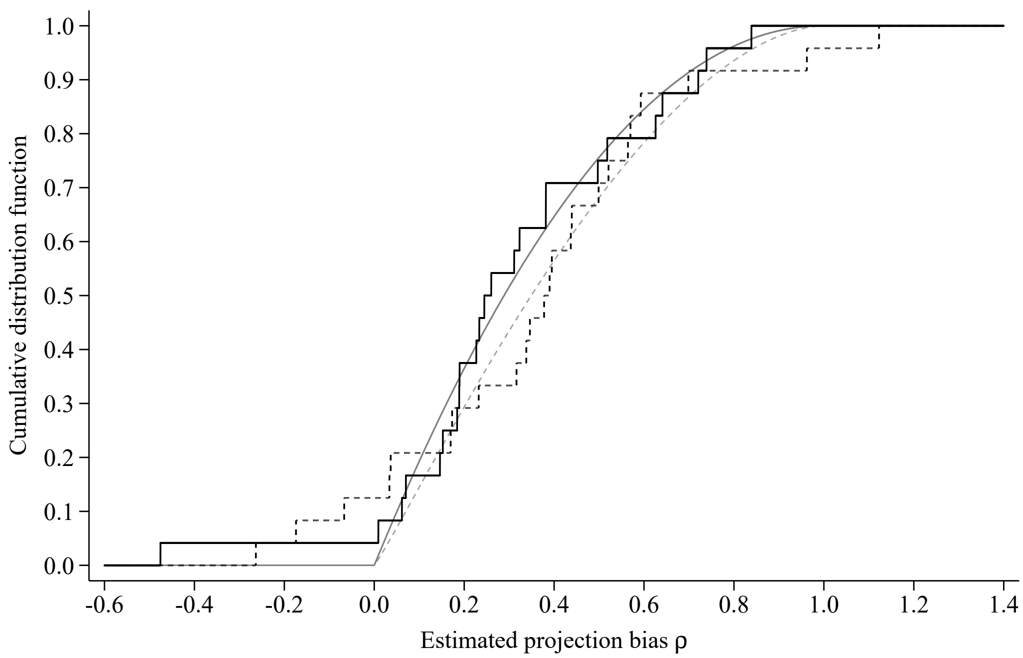


Figure 13: CDFs of principals' (solid) and agents' (dashed) projection bias ρ in the informed treatment with alternative specification replacing the success rates π_t with the agents' first-order estimates $b^I_{A_jt}$ in (4). Black lines represent empirical CDFs; gray lines represent best-fitting beta CDFs.