DISCUSSION PAPER SERIES

DP12867

THE BIASES OF OTHERS: PROJECTION EQUILIBRIUM IN AN AGENCY SETTING

David Danz, Kristóf Madarász and Stephanie Wang

INDUSTRIAL ORGANIZATION

ISSN 0265-8003

THE BIASES OF OTHERS: PROJECTION EQUILIBRIUM IN AN AGENCY SETTING

David Danz, Kristóf Madarász and Stephanie Wang

Discussion Paper DP12867
Published 13 April 2018
Submitted 13 April 2018

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

THE BIASES OF OTHERS: PROJECTION EQUILIBRIUM IN AN AGENCY SETTING

Abstract

We study strategic reasoning and the beliefs people form about the beliefs of others in the presence of private information. We find that while people naively project and think others have the same information as they do, they also anticipate the analogous projection of their differentially-informed opponents onto them. In turn, the typical person explicitly thinks that others form systematically biased beliefs. Specifically, our paper directly tests the model of projection equilibrium, Madarasz (2014, revised 2016), which posits a parsimonious one-to-one relationship between the partial extent to which a player projects and forms biased beliefs about the beliefs others, ?, and the partial extent to which she anticipates but underestimate the same systematic bias in others' beliefs of her beliefs, $?^2$. We find that the distribution of the partial extent to which players project onto others and the distribution of the partial extent to which they anticipate others' projection onto them is remarkably consistent with the tight link implied by the model.

JEL Classification: C9, D2, D8, D9

Keywords: N/A

David Danz - danz@pitt.edu
Pittsburgh and WZB

Kristóf Madarász - k.p.madarasz@lse.ac.uk
London School of Economics and CEPR

Stephanie Wang - swwang@pitt.edu
Pittsburgh

Acknowledgements
We are grateful to Douglas Bernheim, 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: July 2014.

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

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

Current Draft: December 2017.

Abstract

We study strategic reasoning and the beliefs people form about the beliefs of others in the presence of private information. We find that while people naively project and think others have the same information as they do, they also anticipate the analogous projection of their differentially-informed opponents onto them. In turn, the typical person explicitly thinks that others form systematically biased beliefs. Specifically, our paper directly tests the model of projection equilibrium, Madarasz (2014, revised 2016), which posits a parsimonious one-to-one relationship between the partial extent to which a player projects and forms biased beliefs about the beliefs others, ρ, and the partial extent to which she anticipates but underestimates the same systematic bias in others' beliefs of her beliefs, ρ^2. We find that the distribution of the partial extent to which players project onto others and the distribution of the partial extent to which they anticipate others' projection onto them is remarkably consistent with the tight link implied by the model.

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

*We are grateful to Douglas Bernheim, 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: July 2014.

†University of Pittsburgh and WZB Berlin.

‡London School of Economics and Political Science and CEPR

§University of Pittsburgh.

1 Introduction

> "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).

There is growing interest in the behavior of economic agents whose choices are guided by systematically biased beliefs about the environment or their own prospective behavior (e.g., Tversky and Kahneman 1974, Gennaioli and Shleifer 2010, Bénabou and Tirole 2016, Augenblick and Rabin 2017). At the same time, there is little structured evidence about whether people anticipate the same kind of mistakes that color their own judgements in the judgements of *others*, and whether people think and are potentially mistaken about the extent to which others form biased beliefs.

The structure of a individual bias can be studied separately from a person's anticipation of the same kind of misprediction in others. For example in self-control problems existing evidence suggests that while people are naively optimistic about the time they will take to complete a task, they are significantly more pessimistic about the time others will take (Buehler, Griffin, and Ross, 1994).[1] In the context where people have biased beliefs about the beliefs of others, however, that is when the bias is social, studying the structure of a systematic mistake separately from its anticipation in others, is not even possible.

In particular this paper considers the fundamental human capacity, sometimes referred to as theory of mind, of forming beliefs about the beliefs of others. Such

[1] Similarly, in the context of the classic endowment effect research shows that people underestimate the change in their own and others' preferences, Loewenstein and Adler (1995), Van Boven et al. (2003).

beliefs are the building block of social cognition and thus are essential for strategic behavior. Evidence from so-called false belief tasks (Piaget, 1953; Wimmer and Perner, 1983), hindsight tasks (Fischhoff, 1975), curse-of-knowledge tasks (Camerer et al., 1989), the illusion of transparency (Gilovich et al., 1998), and the outcome bias (Baron and Hershey, 1988), point to a robust mistake in this domain.[2] Specifically, in the presence of private information people appear to engage in a form of 'ego-centric' thinking; they project their information onto others and too often assume that others know what they do.

In strategic settings with private information what typically matters, however, is not simply what basic information a person assigns to others, that is, her perception of her opponent's first-order beliefs about the payoff state. Instead what is often equally key is what she thinks others think she believes, that is, her higher-order beliefs. Such views guide her expectation of her opponent's expectation of her beliefs and play an essential role in strategic choice and equilibrium analysis.[3] In turn, even to formulate the psychological phenomenon of projection for strategic settings one must define a person's basic mistake and her anticipation of this mistake in others simultaneously. Through the model of projection equilibrium, Madarasz (2014, revised 2016) provides such a more comprehensive account.

Such a joint account is key for understanding even the most basic psychological implications of this phenomenon. A game theory teacher who projects her information and thus exhibits the curse of knowledge too often thinks that her students should already know what a Nash equilibrium is. Does she, however, anticipate that the students too often expect her to know that they have very little idea about what a Nash equilibrium is? Will this annul or possibly contradict the direct implications of the curse-of-knowledge? A projecting poker player exaggerates the chance that others can read her card. When deciding to bluff does she, however, act as if she exaggerated or underestimated the extent to which others may wrongly think that she can read their card? To what extent does this then

[2] Although hindsight bias is sometimes described as an intrapersonal phenomenon, the evidence is predominantly from interpersonal settings. Evidence also indicates that projection is usually robust to various de-biasing attempts (Wu et al., 2012).

[3] In the context of perfect-information games a number of studies have elicited higher-order beliefs about play, e.g., Bhatt and Camerer (2005) study second-order beliefs and their neural correlates in dominance solvable games; Manski and Neri (2013) study the coherence between first- and second-order beliefs about play in hide-and-seek games.

mean that she thinks that others actually cannot read her mind and do not realize that she does not know their cards?

Such a joint account of the basic mistake and its anticipation in others, and whether people under- or over-exploit each other's judgemental error, also has direct economic implications. In agency settings, a principal — a board, an investor, or a judge — evaluates the quality of an agent — an executive, a financial professional, a public bureaucrat, or a police officer — by monitoring using performance measures that contain ex-post information. An agent who understands that the principal will project such ex-post information anticipates that the principal will underestimate the agent's quality on average, i.e., the principal's updating process will correspond to a super-martingale. In turn, 'to cover her back,' she will want to engage in defensive practices aimed at reducing the informational gap between the ex-ante and the ex-post stages, e.g., distort the production of ex-ante information, undertake an ineffective selection of tasks, or simply dis-invest from an otherwise efficient relationship, (Madarasz, 2012).[4]

The consequences of such anticipation are then often more economically relevant than the basic mistake per se. Effective organizational designs must then be tailored both to people's tendency to project and whether they under- or overestimate such tendency in others. More generally, in strategic settings, from bargaining and trade to optimal auctions or coordination in global games, e.g., Morris and Shin (2003), both the extent to which a person projects and the extent to which she simultaneously thinks that others project is of direct importance.

In the absence of further guidance, there is a bewildering variety of ways in which the higher-order implications of projection may be specified. Clearly, if a player has unbiased views about her opponent's belief hierarchy, then, by construction, she herself cannot systematically project. A salient approach may

[4]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."

then be to adopt a sharply dichotomous classification, as is the norm for individual biases, whereby a person is either fully sophisticated or is fully unaware of the biases of others. The latter types may or may not engage in basic projection, i.e., have systematic biased views of their opponents' first-order beliefs. One could also suggest in coherent ways that the more a person projects the less she thinks that others think she has those beliefs, i.e., the more she thinks others have biased views of her second-order beliefs.

A more skeptical stance would be to opt for a flexible approach, with potentially many degrees of freedom, or to simply reject the idea of a portable approach of this phenomenon and a more general relationship between the extent to which the average person is biased in a context and the extent to which she anticipates the biases of others in that context.

In contrast to such a skeptical stance, the model of projection equilibrium, Madarasz (2014, revised 2016), offers a fully specified, portable yet parsimonious account. It proposes a model of partial projection governed by a single scalar $\rho \in [0, 1)$. By proposing a notion of all-encompassing projection, it links the extent of one's mistaken beliefs about other's first-order beliefs to the extent of her mistaken beliefs along the entire belief hierarchy. It postulates that the systematic bias along this hierarchy vanishes in a tight polynomial fashion.

Crucially, the model implies a one-to-one relationship between the extent to which a player has biased views about her opponent's first-order beliefs (*first-degree projection*) and the extent to which she has biased views about her opponent's second-order beliefs (*second-degree projection*), that is, the extent to which she fails to anticipate his projection onto her. Our paper introduces a careful experimental design to understand the economically key issue of anticipation and the extent to which this parsimonious link, and hence the proposed model of projection for strategic settings, may help capture key aspects of the data.

In our experiment, principals estimated the average success rate π of reference agents in a real-effort task. While the agent never received the solution to the task, principals received the solution to the task prior to the estimation in the informed treatment, as in the case of monitoring with ex-post information, but not in the uninformed treatment. Projection equilibrium predicts that a principal in the informed, but not in the uninformed treatment, should systematically overestimate

the agents' success rate on average. We find that in the uninformed treatment principal are well-calibrated and, consistent with previous qualitative results, e.g., Loewenstein et al. (2006), we find a very strong exaggeration in the informed treatment. The average difference between the principals' first-order estimates across the two treatments allows us to identify the extent of first-degree projection.

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 general 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.

We then turn to the main point of our paper. We elicited the agent's first-order and second-order estimates (their estimate of the principals' estimates) of the success rate of the other agents. In the uninformed treatment, projection equilibrium, just as the unbiased BNE, predicts that the agent's first- and second-order estimates, as well as the principal's first-order estimate, should be correct on average. 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 for this equivalence. 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 on average, and (ii) her second-order estimate shall be lower than the principal's first-order estimate on average. We find exactly this pattern. As implied by projection equilibrium players anticipate that others are biased but underestimate its extent. These findings are summarized by Figure 3.

We then consider our key hypothesis. Projection equilibrium predicts that the partial extent of first-degree projection, that is, the extent to which the principal exaggerates the success rate, shall fully pin down the partial extent of second-degree projection, that is, the extent to which the agent under-estimates the principal's exaggeration on average. If the former is $\rho(1-\pi)$ the latter shall be $\rho^2(1-\pi)$ and vice versa. In turn, we compare the empirical model where the mistake in

the principal's estimate and the mistake in the agent's estimate of the principal's estimate are allowed to freely differ with the empirical model where the former is forced to be the square-root of the latter. This leads to our key finding; the parameter estimates and log-likelihoods of the two empirical models are very similar. The data thus lends strong support to the model of projection equilibrium. These findings are summarized by Table 2.

Finally, we describe (i) the distribution of projection inferred from the distribution of first-degree projection in the principal population and (ii) the distribution of the degree projection inferred from the distribution of second-degree projection in the agent population. We document three facts. First, the majority of the principals 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 and underestimate its extent. We find that the majority of people *partially* believe that others are *partially* biased. Third, and perhaps most surprisingly, consistent with the logic of projection equilibrium, we find that these two distributions are remarkably close to each other. These results are summarized by Figure 4.

To the best of our knowledge, our paper is the first to directly and cleanly measure people's biased beliefs about other people's biased beliefs, and do so in a real effort task, 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 change-detection tasks. In each 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 stim-

ulus (Rensink et al., 1997; Simons and Levin, 1997) and has been studied in the context of the curse-of-knowledge, Loewenstein et al. (2006).

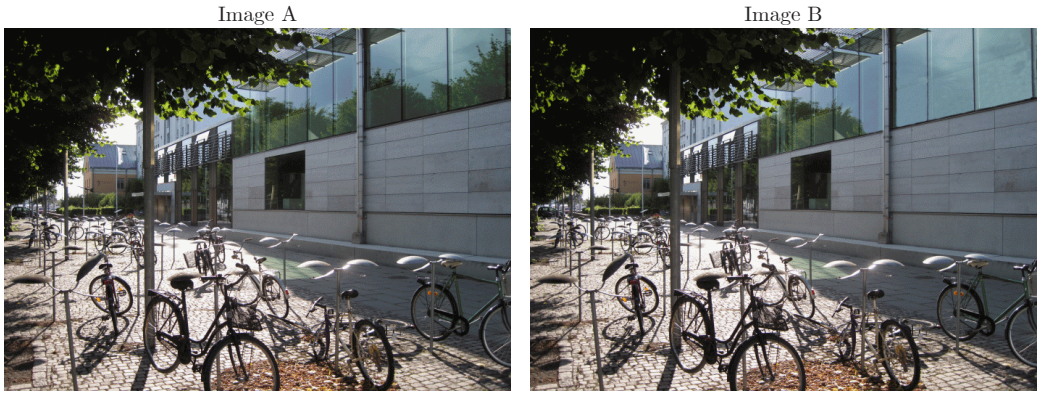


Figure 1: Example of an image pair. Image sequence in the experiment: A, B, A, B, ...

We presented each task in a short clip in which the two images were displayed alternately with short interruptions.[5] Afterwards, subjects could submit an answer by indicating the location of the difference on a grid (see Instructions in the Appendix).[6]

2.2 Principals

Principals had to estimate the performance of others in the change-detection 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 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 20 rounds, the principals were first exposed to the task; that is, they inspected the images and could then submit a solution to the task. Afterwards, the principals stated their estimate (b_t^P) about the fraction of reference agents

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

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

who spotted the difference in that task (success rate π_t henceforth). After each principal stated his or her belief, the next round started.[7]

For the principals the two treatments differed as follows. In the *informed (*asymmetric*) treatment*, the principals received the solution to each 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 a tragic accident, a realized portfolio allocation decision, or a medical case, where principals first learn the outcome and then review the case solved by the agent.

In the *uninformed (*symmetric*) treatment*, the principals were not given solutions to the tasks (the same image was shown on the countdown screen, but without the red circle and the corresponding note in Figure 2). The principals did not receive feedback during the experiment. Principals in both treatments then went through each task exactly as the reference agents did.

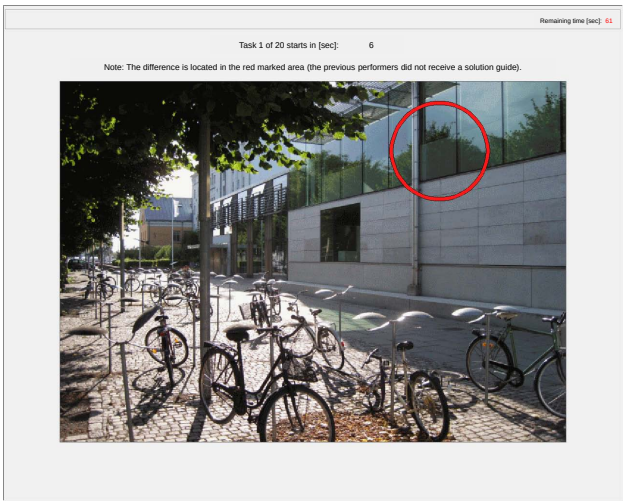


Figure 2: Screenshot from the treatment with informed principals: Countdown to the next task providing the solution (translated from German).

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

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 $b_t^P \in [\pi_t - 0.05, \pi_t + 0.05]$, 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.

2.3 Agents

Agents in both treatments were informed that in previous sessions (i) reference agents had performed the change-detection tasks being paid according to their performance and (ii) principals had estimated the average performance of the reference agents being paid according to the accuracy of their estimates.[8]

The agents where 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 only with respect to the kind of principal they where matched to: In the informed treatment, agents were randomly matched to one of the informed principals; in the uninformed treatment, agents where 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.

In each of 20 rounds, the agents in both treatments first performed the task in the same way as the reference agents; that is, they went through the images and then submitted an answer. Next, agents received feedback regarding the solution to each 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 this feedback corresponded to what the principal had seen for that task. Agents matched to uninformed principals were told the principals had not received the solution to the task. In both treatments, the agents did not receive information about the principal's estimates.

[8]This aspect of the design prevents the possibility of any collusion equilibria where principals and agents co-ordinated on sub-optimal task performance.

In some session, in addition the agents' decided between two options. One of the options provided a sure payoff of €4. The payoff of the other option depended on the success rate of the reference agents, π_t, and the corresponding estimate of the principal matched to her. Specifically, the agent received €10 if the principal's estimate b_t^P was not more than 10 percentage points higher than the success rate π_t; otherwise, the payoff was €0. Choosing this option can be thought of an investment whose expected return is decreasing in the principals' expectations of the agents' likelihood of success. Similarly, choosing the sure payoff can be thought of buying insurance against overly optimistic expectations of the principals. Throughout the paper, we will refer to this choice as the agents' investment decision.[9]

In other sessions belief elicitation replaced the investment decisions for the first ten tasks of the experiment. Specifically, following each of the first 10 change-detection tasks the agents stated (i) their estimate about the fraction of reference agents that spotted the difference in that task (*first-order estimate* $b_{1,t}^A$ henceforth) and (ii) their estimate about their principal's estimate of that success rate (*second-order estimate* $b_{2,t}^A$ henceforth).

The 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 beliefs for payment, namely, either her first- or second-order belief in that round.[10] The subject received €12 if her stated belief was within five percentage points of the actual value (the actual success rate in case of a first-order belief and the principal's estimate of that success rate

[9] In terms of the introductory example of defensive medicine, the principal in the informed treatment can be thought of a judge or juror who, in hindsight, forms expectations about the radiologist's ex ante chance of correctly diagnosing a patient based on a radiograph. At the time of her evaluation, the juror has access to both the initial radiograph and further information about the patient's condition revealed later on. The biased juror projects this data on the radiologist's information set when diagnosing the patient and consequently overestimates her ex ante chances of success. In turn, the physician who anticipates this bias will exhibit a higher willingness to pay for insurance against malpractice claims or simply does not invest in career paths that are particularly susceptible to biased ex post performance evaluations.

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

Table 1: Overview of treatments and sessions.

	Treatment	
	Informed	**Uninformed**
Principals	Elicitation of first-order beliefs regarding the success rate of reference agents in 20 change-detection tasks	
Information	Get solutions to change-detection tasks	Do not get solutions to change-detection tasks
Sessions	1	1
Subjects	24	24
Agents	Choices between a sure payoff and investment whose expected return is decreasing in principals' expectations about the success rate; *Elicitation of first-order beliefs (estimate of the success rate) and second-order beliefs (estimates of the principals' estimate)	
Information	Know that principals have access to solutions	Know that principals do not have access to solutions
Sessions	3	3
Subjects	48 $(12 + 12 + 24^*)$	46 $(11 + 12 + 23^*)$

Note: Asterisks indicate sessions with belief elicitation instead of investment decisions in the first half of the experiment.

in case of a second-order belief), and nothing otherwise.[11] If the round selected for payment involved an investment decision, the agent was paid according to her decision. Table 1 provides an overview of the treatments and sessions.

Since agents solve each task in both treatments without any feedback, when deriving the implications of projection equilibrium below we consistently assume that agents project vis-a-vis the informational conditions of the task when being solved. This assumption is validated by the data on agents' first-order estimates—

[11] 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.

both conditionally on task performance and on average— are identical across the two treatments.

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.[12] 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.[13] 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

Our predictions are based on the model of projection equilibrium introduced in Madarasz (2014, revised 2016).[14] This model pins down the meaning of information projection for strategic settings by specifying what each player i (mistakenly) believes about her opponents' belief hierarchies. A key aspect of the model is the parsimonious way it ties together the extent to which people are biased and the extent to which they anticipate the biases of others.

To describe the predictions for our design, fix a basic task. Let there be a finite space 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 $P_i : \Omega \to 2^{\Omega}$. Player i's action set is A_i and her payoff is $u_i(\omega, a)$ where $a \in A = \Pi_{i=1}^{N} A_i$. In our design, there are only two strategically active players: one agent and one

[12] 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.

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

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

principal. The reference agents are strategically passive; they perform only the basic task and their payoffs do not depend on the choices of the other players. In what follows subscript A then refers to the strategically active agent and subscript P to the principal. The game can be summarized by $\Gamma = \{\Omega, \phi, P_i, A_i, u_i, N\}$.

Solving the basic task is equivalent to picking a cell $x \in D$ from the finite grid on the visual image. This is performed by all players. The action set of the principal also includes his estimation task, hence, $A_P = D \times [0,1]$. The action set of the strategically active agent also includes her two estimation tasks, hence, $A_A = D \times [0,1] \times [0,1]$. In our design no player i's payoff from choosing x_i interacts with her payoff from the potential other decisions.[15] Hence, we can separate the payoff from the basic task, denote it by $f(x_i, \omega)$, and normalize it to be one if the solution is a success, and zero otherwise.

Projection Equilibrium. Under projection equilibrium, each player i best responds to a biased perception of each of her opponent's strategy: she attaches probability $1-\rho_i$ to player j's true strategy and probability ρ_i to a strategy played by a fictional projected version of j. This projected version of j conditions his behavior on the exact same information about the payoff state as player i does.[16] In equilibrium, this projected version of j, given such beliefs, best responds to i's true strategy.

The model implicitly gives rise to a structure of systematically biased beliefs along each player's belief hierarchy. The key assumption is that projection is all-encompassing; in each state, player i believes that the projected version of j assigns probability one to i's true type, that is, her actual belief hierarchy. This structure pins down the predictions for our design. Ingrid believes not simply that projected Jack has the exact same belief about the solution as she does, i.e., has the same first-order belief about the solution. Ingrid also believes that projected Jack knows what she thinks about the solution, i.e., that his second-order belief assigns probability one to Ingrid's actual first-order belief. Similarly, she thinks that projected Jack knows her actual second-order belief and so on. As

[15] Note that the strategically active players always estimate the success rate of the strategically passive agents. Hence, this separation then ensures that there cannot be an equilibrium where agents and principals may co-ordinate on sub-optimal performance on the basic task.

[16] In turn, under projection equilibrium people project both their information and their ignorance.

the predictions below will highlight, this then leads to a polynomially vanishing bias structure in higher-order beliefs.

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, $\rho_P \neq \rho_A$ may hold. This specification will still greatly restrict the set of possible outcomes in our design.

Homogeneous Projection. The case of homogeneous projection, $\rho_A = \rho_P$, is of particular interest for our design. It implies a one-to-one relationship between the principal's first-degree projection and the agent's second-degree projection. Specifically, it implies that the extent of first-degree projection, as measured by the systematic wedge between the principal's first-order estimate and the truth, perfectly pins down the extent of second-degree projection, as measured by the wedge between the agent's second-order estimate and the principal's first-order estimate.

Before turning to the predictions, three additional remarks are in order. First, in our design, any two players i and j may obtain different private signal realizations from watching the two images. 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. Second, the predictions below hold in the ex-ante expected sense. We do not focus on conditional estimates, that is, estimates conditional on a player's performance on the basic task. Instead we focus on the average estimates within a treatment and the predictions below express differences about such average estimates across treatments. In Section 5, we return to the relevant issue of conditional estimates and projection. Finally, the predictions below nest the unbiased BNE as a special case. Here, $\rho_i = 0$ for $i \in \{P, A\}$.

Consider first the ex-ante probability with which a randomly chosen player i who sees only the two images can solve the 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 probability of successfully solving 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 treatment both the agent and the principal have access to the two images only. Hence, by the law of iterated expectations, $d = 0$ must hold here. In the informed treatment, by contrast, the principal also has access to the solution; since the solution always helps solve the basic task, $d \simeq 1 - \pi > 0$ must hold in the informed treatment.

Claim 1. *Under projection equilibrium the principal's ex-ante expected estimate of the success rate is given by* $\pi + \rho_P d$.

In the uninformed treatment, the principal's estimate is unbiased. While projection will distort conditional beliefs, that is, the principal's estimate of the success rate conditional on her own success or failure on the basic task may well be distorted, such distortions must cancel out on average. Since roles are determined randomly, the agent and the principal have the same ex-ante probability of success on the basic task. Even if the principal fully projects, hence, predicts success whenever she figures out the solution and likely failure whenever she does not, her estimate is correct *on average*. The same holds for all ρ_P.

In the informed treatment in contrast, the principal exaggerates the success rate on average. Since she always knows the solution, by projecting, she exaggerates the probability with which the agent figures out the solution. The ex-ante expected exaggeration amounts to $\rho_P d$.

We now turn to the key hypothesis of our paper. The model implies a systematic wedge between an agent's second-order estimate, her estimate of the principal's estimate of the success rate, and her *own* first-order estimate of the success rate, in the informed treatment, but not in the uninformed treatment, on average.

Claim 2. *Under projection equilibrium:*

1. The agent's ex-ante expected first-order estimate of the success rate is π.

2. The agent's ex-ante expected second-order estimate of the success rate is $\pi + (1 - \rho_A)\rho_P d$.

In the uninformed treatment, the agent's first-order estimate is again unbiased. The reason is exactly the same as for the principal's estimate in the uninformed treatment. There is also no systematic wedge between either the agent's second-order estimate and her own first-order estimate or between the agent's second-order estimate and the principal's first-order estimate. All these equal to π on average.

In the informed treatment, the agent's first-order estimate is still unbiased, but there are two systematic departures from the predictions of the unbiased BNE; (i) the agent's second-order estimate shall be systematically *higher* than her own first-order estimate and (ii) her 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 her own projection, she underestimates the principal's projection.[17]

To describe the logic, note that if $\rho_A = 0$, only the first point holds. An unbiased agent fully anticipates the principal's bias on average. In contrast, if $\rho_A \to 1$, only the second point holds. A fully biased agent thinks that the principal always has the same beliefs as she does. Given any partially biased agent, both of the above statements hold strictly. Under heterogenous projection the agent always *anticipates* and *underestimates* the principal's exaggeration on average.

Under homogeneous projection the extent of the principal exaggeration, ρd, directly pins down the extent of the agent's underestimation of this exaggeration, $\rho^2 d$. The table below then summarizes the tight ex ante expected polynomially vanishing bias structure along the players' belief hierarchies:

Ex-ante expected bias	Uninformed	Informed
bias in the principal's first-order estimate	0	$\rho(1-\pi)$
bias in the agent's first-order estimate	0	0
bias in the agent's second-order estimate	0	$-\rho^2(1-\pi)$.

[17] Under projection equilibrium player i thinks that her projected opponents occur in a perfectly correlated manner. Furthermore, these projected opponents also believe this.

Finally, we also consider the setting where the principal's action set is unchanged, but where the strategically agent's action set is, $A_A = D \times \{\text{Invest, Not Invest}\}$, that is, she decides whether to opt for the sure payment or the lottery whose expected return is decreasing in the principal's first-order estimate.

Claim 3. *The agent's propensity to invest is lower in the informed than in the uninformed treatment on average.*

Under projection equilibrium the above comparative static holds for any $\rho_P, \rho_A \in (0, 1)$. Since the agent has the same beliefs about the strategy of the projected version of the principal in both treatments and assigns positive weight to the principal's true strategy, thus partially anticipates the principal's exaggeration in the informed treatment, she is predicted to invest more often in the uninformed than in the informed treatment.

4 Results

4.1 Principals

The principals' estimates support 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% ($p = 0.824$).[18] In contrast, principals in the informed treatment grossly overestimate the success rate for the vast majority of tasks.[19] 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 ($p < 0.001$). Principals in the informed treatment, also then had significantly lower

[18] 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 statistics per subject.

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

expected earnings (€2.40) than principals in the unformed treatment (€3.65; one-sided t-test: $p = 0.034$).[20]

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

The systematically biased forecasts in Result 1 are not surprising given previous findings in the literature and it also confirms the basic premise of our design.[21]

4.2 Agents

4.2.1 Investment decisions

We now move to the agent's anticipation (or potential anti-anticipation) of the principal's bias. Consider first the investment decisions. Recall that, for the agents, the only difference between the two treatments is that agents in the informed were told that their principal had access to the solution while agents in the uninformed treatment were told that their principal had not seen the solution.

The data clearly supports Claim 3.[22] Agents matched to informed principals invest at a significantly lower rate than agents matched to uninformed principals.[23,24] The average investment rate of agents matched to uninformed principals

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

[21] 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.

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

[23] 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.

[24] 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

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.*

Result 2 is consistent with agents anticipating the projection of the principals. 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 estimate.

4.2.2 Stated Beliefs

We now turn to the key hypothesis of our study and consider the agent's first- and second-order estimates of the success rate. Figure 3 summarizes our key findings. It shows a bar chart of the average stated estimates of the agents in each treatment together with the true success rate over all tasks and the corresponding estimates of the principals.

The left panel shows the data in the uninformed treatment. Under projection equilibrium, all stated estimates shall be correct on average. This is indeed what we find. None of the elicited beliefs are, on average, significantly different from the true success rate: neither the agents' first-order estimates ($p = 0.917$), their second-order estimates ($p = 0.140$), nor the principals' estimates ($p = 0.337$). Although the agents' second-order estimates are somewhat higher than the true estimates of the principals, this difference is not significant either ($p = 0.080$).

The right panel shows the data in the informed treatments. While the principals vastly overestimate the true success rate, the agents are well calibrated ($p = 0.967$) on average. Crucially, as predicted, the agents' second-order estimates are also significantly higher than their *own* first-order estimates ($p < 0.001$) on average. Finally, as predicted, their second-order estimates are also significantly lower than the principals' estimates (one-sided t -test: $p = 0.047$). Exactly as pre-

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).

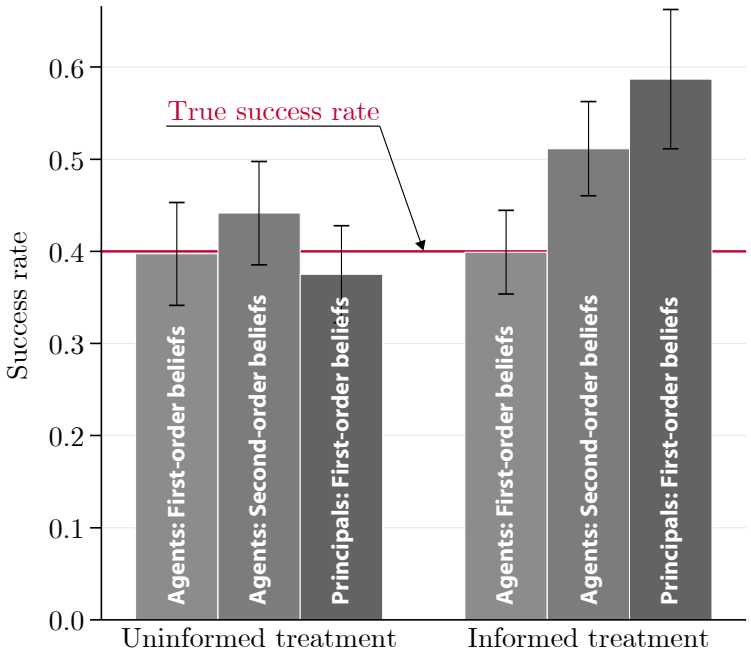


Figure 3: Agents' first-order estimates (estimates of the success rate) and second-order estimates (estimates of the principals' estimate), conditional on being matched with informed or uninformed principals. Capped spikes represent 95% confidence intervals.

dicted, agents appear to both anticipate and underestimate the systematic mistake in the principals' estimates.

When comparing treatments, the agents' first-order estimates are not significantly different ($p = 0.956$) across treatments. The average second-order estimate of agents in the informed treatment 51.14% is, however, significantly higher than that in the uninformed treatment 41.15% ($p = 0.0314$ for one-sided t-test). Finally, the average difference between the agents' second- and own first-order estimates is also significantly larger in the informed than in the uninformed treatment ($p < 0.001$).[25] In sum,

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

[25]This treatment difference is robust to controlling for individual characteristics (see Table 7 in the Appendix).

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 BNE but confirms all predictions of projection equilibrium under heterogeneous projection.

4.3 Estimation of Projection Equilibrium

The analysis has confirmed all four predictions of projection equilibrium under heterogeneous projection. A key aspect of projection equilibrium and its parsimony, as mentioned, however, is the fact that the extent of a player's first-degree projection fully pins down the extent of her second-degree projection. Yet 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 these two can then differ substantially. To test this key quantitative aspect of the model, we then have to consider homogeneous projection, i.e., $\rho = \rho_A = \rho_P$.

Under homogeneous projection, the extent to which the principal exaggerates the success rate in the informed treatment, $\rho(1-\pi)$, should fully determine the extent to which the agent underestimates this exaggeration in this treatment, $\rho^2(1-\pi)$. Claims 1 and 2 then provide two separate equations for measuring ρ. Based on the average estimates of the players as described in the previous section, using Claim 1 we obtain the estimate of $\widehat{\rho} = 0.3$ while using Claim 2 we obtain the estimate of $\widehat{\rho} = 0.31$. Similarly, estimating Claims 1 and 2 under heterogeneous projection we obtain that the pair $\widehat{\rho}_P = 0.3$ and $\widehat{\rho}_A = 0.33$ provides the unique solution to these equations.[26]

[26]Solving Claim 1 under homogeneous projection, given the fact that $\pi = 0.39$,

$$0.57 - \pi = (1-\pi)\rho$$

leads to $\widehat{\rho} = 0.3$ and solving Claim 2 under homogeneous projection

The above estimates points to the key finding of our paper. Remarkably, the ratio between the agent's underestimation of the principal's exaggeration over the principal's exaggeration is very close to the principal's exaggeration divided by $1-\pi$. As predicted by projection equilibrium, the extent to which the typical player projects onto others, the first-degree projection, matches the extent to which the typical player underestimates other's projection onto her, the second-degree projection.

4.3.1 An Econometric Analysis

To examine the above finding further, we conduct an econometric analysis. We start with a flexible specification that allows for different degrees of projection both between roles as well as within roles. We denote the average degree of projection in the principal population by ρ_μ^P and that in the agent population by ρ_μ^A. We then estimate the model under the assumption that these averages are the same, i.e., $\rho_\mu^P = \rho_\mu^A$. In our design this provides the ultimate test of the model since it ties directly together the extent of the basic mistake with the extent of the mistake in the estimate of the biases of others.

Following Claim 1, principal i's expected stated estimate of the success rate π_t in task t, conditional on bias ρ_i^P, is:

$$E(b_{1,i,t}^P \mid \pi_t, \rho_i^P) = \pi_t + \rho_i^P(1-\pi_t) \equiv \mu_{it}^P. \tag{1}$$

Following Claim 2, agent j's second-order estimate $b_{2,j,t}^A$, conditional on her first-order estimate and her bias ρ_j^A, is:

$$E(b_{2,j,t}^A \mid b_{1,j,t}^A, \rho_j^A) = b_{1,j,t}^A + (1-\rho_j^A)\rho_\mu^P(1-b_{1,j,t}^A) \equiv \mu_{j,t}^A. \tag{2}$$

$$0.57 - 0.51 = (1-0.39)\rho^2$$

leads to $\widehat{\rho} = 0.31$. Similarly, given Claim 1,

$$0.57 - \pi = (1-\pi)\rho_P$$

is solved by $\rho_P = 0.3$, and then, given Claim 2,

$$0.57 - 0.51 = (1-\pi)\rho_P\rho_A$$

is solved by $\rho_A = 0.33$.

To account for the fact that subjects' stated beliefs fall in a bounded interval between zero and one, we employ a beta regression specification. In particular, we assume that subjects' realized responses $b^P_{1,i,t}$ and $b^A_{2,j,t}$ follow a beta distribution with means μ^P_{it} (1) and $\mu^A_{j,t}$ (2), respectively. To simplify notation, we write b^k_{it} for subject i's stated belief in role k, where $b^k_{it} = b^P_{1,i,t}$ for principals ($k = P$) and $b^k_{it} = b^A_{2,i,t}$ for agents ($k = A$). With beta-distributed responses, the density of stated belief b^k_{it} of subject i in role k can be written as[27]

$$f(b^k_{it}; \mu^k_{it}, \phi_b) = \frac{\Gamma(\phi_b)}{\Gamma(\phi_b \mu^k_{it})\Gamma(\phi_b(1-\mu^k_{it}))}(b^k_{it})^{\phi_b \mu^k_{it}-1}(1-b^k_{it})^{\phi_b(1-\mu^k_{it})-1},$$

where ϕ_b is a precision parameter that is negatively related to the noise in subjects' response.[28]

To capture unobserved individual heterogeneity in the projection parameters and to account for repeated observations on the individual level, we employ a random coefficients model. Specifically, we assume that the individual-specific degree of projection $\rho^k_i \in [0,1], k \in \{A, P\}$ follows a beta distribution with role-specific mean ρ^k_μ and density

$$g(\rho^k_i; \rho^k_\mu, \phi_\rho) = \frac{\Gamma(\phi_\rho)}{\Gamma(\phi_\rho \rho^k_\mu)\Gamma(\phi_\rho(1-\rho^k_\mu))}(\rho^k_i)^{\phi_\rho \rho^k_\mu-1}(1-\rho^k_i)^{\phi_\rho(1-\rho^k_\mu)-1},$$

where the parameter ϕ_ρ is negatively related to the variance of projection bias in the principal and the agent population.[29]

We now formulate the log-likelihood function. Conditional on ρ^k_i and ϕ_ρ, the likelihood of observing the sequence of stated beliefs $(b^k_{it})_t$ of subject i in role k is

[27] See Ferrari and Cribari-Neto (2004) for this convenient parametrization. A link function as in standard beta regression models is not needed because (1) and (2) map $(0,1) \longrightarrow (0,1)$ for $\rho \in [0,1)$ and $\pi_t, b^A_{1,i,t} \in (0,1)$. Abstaining from using a link function has the advantage that we can interpret the estimated parameters ρ^P_μ and ρ^A_μ directly as parameters of projection from Claims 1 and 2. As in standard beta regression models, observations of $y^k_{it} = 0$ or $y^k_{it} = 1$ have a likelihood of 0 and can therefore not be used in maximum likelihood estimation. However, the entire data set has only one observation of $y^k_{it} = 1$ (one of the principals' first-order estimates is $b^P_{1,i,t} = 1$). We drop this observation in the estimation. Treating this observation as $\acute{b}^P_{1,i,t} = b^P_{1,i,t} - \varepsilon, \varepsilon = 10\hat{}(-10)$ yields similar results.

[28] The variance of b^k_{it} is given by $\mu^k_{it}(1-\mu^k_{it})/(1+\phi_b)$, i.e., conditional on the expected belief statement μ^k_{it}, noise in response is decreasing in ϕ_b (see Ferrari and Cribari-Neto, 2004).

[29] The variance of ρ^k_i is given by $\rho^k_\mu(1-\rho^k_\mu)/(1+\phi_\rho)$.

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

	Unrestricted model with heterogeneous ρ ($\rho_\mu^P \neq \rho_\mu^A$)		Restricted model with homogeneous ρ ($\rho_\mu^P = \rho_\mu^A$)	
Parameter	Estimate	Conf. interval	Estimate	Conf. interval
ρ_μ^P	0.279^{***}	$[0.198, 0.359]$	0.274^{***}	$[0.211, 0.338]$
ρ_μ^A	0.299^{**}	$[0.036, 0.562]$		
ϕ_ρ	4.013^{***}	$[1.651, 6.375]$	3.965^{***}	$[1.673, 6.257]$
ϕ_b	5.156^{***}	$[4.637, 5.675]$	5.155^{***}	$[4.637, 5.674]$
N	720		720	
$\ln L$	184.917		184.898	

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

given by

$$L_i^k(\rho_i^k, \phi_b) = \prod_t f(b_{it}^k; \mu_{it}^k(\rho_i^k), \phi_b).$$

Hence, the unconditional probability amounts to

$$L_i^k(\rho_\mu^k, \phi_\rho, \phi_b) = \int \left[\prod_t f(b_{it}^k; \mu_{it}^k(\rho_i^k), \phi_b)\right] g(\rho_i^k; \rho_\mu^k, \phi_\rho) d\rho_i^k. \tag{3}$$

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

$$\ln L(\rho_\mu^P, \rho_\mu^A, \phi_\rho, \phi_b) = \sum_k \sum_i \log L_i^k(\rho_\mu^k, \phi_\rho, \phi_b). \tag{4}$$

We estimate the parameters in (4) simultaneously by maximum simulated likelihood (Train 2009; Wooldridge, 2010).[30]

[30]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).

Table 2 shows the estimation results for the unrestricted model with $\rho_\mu^P \neq \rho_\mu^A$ in the left column and the restricted model with $\rho_\mu = \rho_\mu^P = \rho_\mu^A$ in the right column.[31] We focus on the unrestricted model first where we make three observations. First, the principals' average degree of projection is estimated to be 0.274 with a confidence interval of $[0.20, 0.36]$. This estimate clearly indicates the relevance of 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 on average, is estimated to be 0.299 with a confidence interval of $[0.04, 0.56]$.[32] The $\hat{\rho}_\mu^A = 0.299$ estimate — which is significantly different from 0 and clearly different from 1 — gives structure to our observation that agents do anticipate that principals are partially biased — but under-anticipate the principals' level of projection.

Crucially, the estimated parameters of projection are not significantly different between the principals and the agents ($p = 0.839$). Furthermore, the log likelihood of the two models are very close, and standard model selection criteria (e.g., BIC) clearly favor the single-parameter model of homogeneous projection (see right columns of Table 2) over the unrestricted model with two parameters.

In short, the data is remarkably consistent with the tight link between the mistake in first-order and second-order beliefs about one's opponent strategy implied by projection equilibrium. The empirical extent of the basic curse-of-knowledge mistake and the empirical extent of the mistake in the anticipation of this basic mistake in others matches the predictions of the model.

[31] 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 (4) appears to assume a global (and unique) maximum at the estimated parameters.

[32] We also did the estimation separately for tasks below median difficulty and tasks above median difficulty. The parameters in these two estimations are similar to each other and to the pooled estimation. Furthermore, Vuong's (1989) test of the model with homogeneous projection in Table 2 against the standard model without projection yields $p < 0.0001$. For the standard model specification, we assume b_{it}^k is beta distributed with mean $\mu_{it}^k = \pi_t$ for principals and $\mu_{it}^k = b_{1,t,j}^A$ for agents, i.e, $b_{it}^k \sim Beta(\mu_{it}^k, \phi_b)$. The estimated precision parameter for the standard model amounts to $\hat{\phi}_b = 2.580$ (s.e. 0.115); the log likelihood of the standard model is 115.801.

4.4 Individual heterogeneity and Partial Projection

The final part of the analysis is devoted to a test of the hypothesis of *partial* projection and partial anticipation at the individual level. This is done in conjunction with a test of the econometric specification of the empirical model (4) above. Specifically, we test whether the mean projection bias $\hat{\rho}_\mu = 0.274$ estimated from (4) is indeed generated by a beta distribution of ρ_i. A misspecification in this matter would not only be relevant from an econometric point of view; it would also challenge the interpretation of our results. Specifically, if the estimate of the mean projection bias was a result of a finite mixture of some agents not anticipating information projection at all ($\rho = 1$) and the remaining agents fully anticipating information projection ($\rho = 0$), then the model (4) would be misspecified and, more importantly, Claim 2 (partial anticipation) would have little empirical bite.

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 the informational projection bias parameter ρ for each principal and each agent from the informed treatment using simple linear regressions without imposing any restrictions on the size or even the sign of the parameters. Specifically, for each informed principal i, we estimate his degree of projection ρ_i^P via

$$b_{1,i,t}^P = \pi_t + \rho_i^P(1 - \pi_t) + \epsilon_{it}, \tag{5}$$

where $b_{1,i,t}^P$ denotes the principal's expectation of the success rate π_t in task t, and ϵ_{it} denotes an independent and normally distributed error term with mean zero and variance σ_i^2. Analogously, for each agent j in the informed treatment we estimate her degree of projection ρ_j^A via

$$b_{2,j,t}^A = b_{1,j,t}^A + (1 - \rho_j^A)\rho_\mu^P(1 - b_{1,j,t}^A) + \epsilon_{jt}, \tag{6}$$

where ρ_μ^P is the mean projection bias in the principal population, and ϵ_{jt} again denotes an independent and normally distributed error term with mean zero and variance σ_j^2. We estimate the parameters in (5) and (6) by OLS, where we sub-

stitute ρ_μ^P in (6) with the average estimate of ρ_i^P obtained from the regressions in (5).[33]

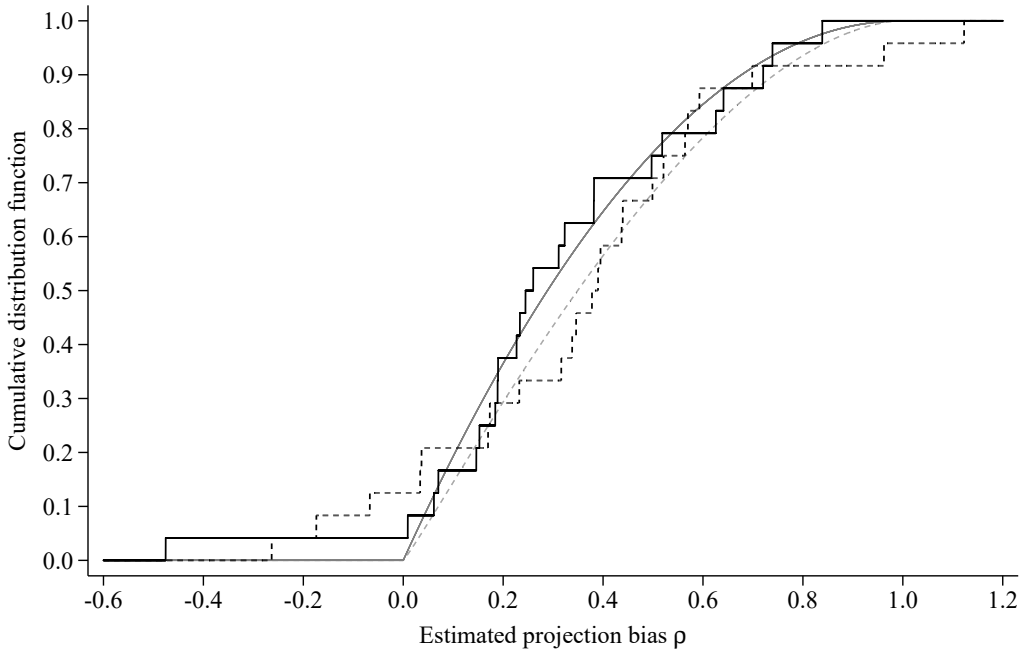


Figure 4: Cumulative distribution functions (CDF) of principals' (solid) and agents' (dashed) projection bias ρ in the informed treatment. Black lines represent empirical CDFs; gray lines represent best-fitting beta CDFs.

Figure 4 shows the empirical CDFs of the individual degrees of projection in the principal and the agent populations. Crucially, a casual inspection of the figure already suggests that the empirical CDFs of the principals' and the agents' ρ 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 the average projection by the principals and the agents are the same, but the two distributions

[33]Because we use an estimate of ρ_μ^P in the estimation of the agents' projection, the composite error term in equation (6) is heteroskedastic. We therefore 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 (4), 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.

are also not significantly different. These findings lend strong support for the key feature of projection equilibrium—that the extent of first-degree projection and the extent of its anticipation, the extent of the second-degree projection, are consistent with the theory and the proposition that these are governed by the same force.

Figure 4 conveys an additional result regarding the econometric specification in (4). The empirical CDFs of principals' and the agents' projection bias are very well described by beta distributions (the gray lines in the figure show the best-fitting beta distributions).[34] Moreover, the joint empirical CDF of the principals' and the agents' projection is not significantly different from the beta distribution $g(\rho_\mu = 0.274, \phi_\rho = 3.965)$ obtained from model (4) with homogeneous projection (see Table 2; Kolmogorov-Smirnov test: $p = 0.063$).

Finally, 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). Players think that others are partially biased and underestimate its extent. 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_j^A that is significantly larger than zero, but significantly smaller than one.[35] People do believe that others are partially biased and partially underestimate its extent.[36]

5 Discussion

Below we discuss the predictions of some alternative models in the context of our design and also the issue of conditional estimates.

[34] Kolmogorov-Smirnov tests of the empirical CDF against the best-fitting beta distribution (as shown in Figure 4) yields $p = 0.941$ for the principals and $p = 0.584$ for the agents.

[35] The second most common category (37.5%) 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. The remaining agents (12.5%) 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. 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: $p = 0.124$).

[36] Section 6.3 in the Appendix shows that the projection equilibrium also outperforms unbiased BNE in out-of-sample predictive accuracy.

5.1 Alternative Models and Mechanisms

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.

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 rather than misperceiving the relationship between other players' beliefs and their actions. In particular, the models of ABEE (Jehiel, 2005; Jehiel and Koessler, 2008), and cursed equilibrium (Eyster and Rabin, 2005), assume that people have coarse or misspecified expectations about the link between others' actions and their information. Crucially, these models are 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.[37]

Risk Aversion. We find no evidence that risk aversion matters for the subjects' choices. (See Tables 4 and 7 in the Appendix). Note that since more information helps unbiased principals make more accurate forecasts on average, under correct beliefs 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 opposite.

Overconfidence. Finally, 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 these estimates are in fact unbiased empirically. Similarly, a principal may be

[37] 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.

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 per se.

5.2 Projection and Conditional Estimates

As mentioned, to avoid selection effects, the need to make further assumptions about the data generating process, and to have a clean contrast with the predictions of unbiased BNE, we compare average beliefs across treatments. At the same time, players' conditional beliefs within a treatment, that is, their estimates conditional on whether or not they themselves were able to solve the task, shall also be affected by projection. In particular, within the uninformed treatment, a player who figures out the solution herself should project this information and may exaggerate the success rate and a player who does not figure it out should project her ignorance and may underestimate the success rate. Indeed we can check whether or not this holds in our data. Consistent with these predictions, principals within the uninformed treatment who spotted the difference overestimated the success rate (60.93%) while those who did not figure out the solution underestimated the same (29.95%).

Note that in contrast to average beliefs, without further assumptions we can not pin down the nature of unbiased conditional estimates. Hence, we cannot identify the bias in conditional beliefs within the uninformed treatment either. At the same time, we can still describe how the estimates of the principals in the informed treatment who were given the solution compare to the estimates of the principals in the uninformed treatment who were not given the solution but who did figure this out themselves. 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 further provides broad support for our premise and 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.

Finally, 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 (see Table 6 in the Appendix).

6 Conclusion

This paper shows that people believe that others have systematically false beliefs. While a host of robust findings demonstrate that people engage in limited informational perspective taking the very meaning and implications of such a phenomenon crucially depend on the extent to which people simultaneously anticipate this tendency in each other. Our study lends surprisingly strong empirical support to the model of projection equilibrium which postulates a tight link between these two. People believe that others are partially biased but, in proportion to their own projection onto them, underestimate its extent.

Providing empirical support for such a joint parsimonious account greatly facilitates the study of this phenomenon in many classic economic problems such as as the relationship between information and incentives in agency problems (e.g., Holmström 1979, Baker et al. 1994), the link between risk taking and the way liability is established in courts (e.g., Harley 2007 argues that hindsight bias is a key factor in the judgement of jurors), or the allocation of authority in organizations (Aghion and Tirole, 1997). In all these settings both the basic mistake and its anticipation in others will 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). The model is able to capture the systematic departure from the predictions of BNE due to the fact that it predicts

both a systematically biased view of the information of others and the anticipation of this mistake in others.[38]

Our results also illustrate the potential of obtaining key insights by eliciting higher-order beliefs in social cognition. While there are multiple factors that might generate biased beliefs in social settings, it is the pattern in higher-order beliefs, beliefs about those biases, that may be essential. 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 predicted by projection equilibrium and established by our findings. Exploring such issues in other contexts as well (e.g., Mobius et al. 2014) may help better understand the nature and the economic implications of biased social cognition.

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 (2017): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." mimeo.

[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.

[38] 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.

[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. (2005): "Analogy-based expectation equilibrium." *Journal of Economic Theory*: 123(2), 81–104.

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

[24] 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.

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

[26] 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/.

[27] 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.

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

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

[30] 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

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

[32] 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.

[33] 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.

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

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

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

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

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

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

[40] 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.

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

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

[43] 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.

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

[45] 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

6.1 Proofs

For the definition of projection equilibrium for N-player games see Madarasz (2016). Since the reference agents solving the task have no direct strategic interaction with each other, we can introduce a representative reference agent and denote it by $\overline{A}$. This is short-hand for the ex-ante expected average performance of the population of 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 $\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)]]$. We will then equate the ex-ante expectation of this with the actual success rate in the population π.

Proof of Claim 1. Let E^{ρ_P} denote the expectation of a ρ_P-biased principal. The ex-ante expected mean estimate of π by a principal is

$$E_\omega \left[E^{\rho_P} [\max_{x \in D} E[f(\omega, d) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \right]. \tag{7}$$

Using the definition of projection equilibrium for a principal with information $P_P(\omega)$, we obtain that

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

Given the law of iterated expectations and the linearity of expectations this then becomes

$$\rho_P E_\omega \left[\max_{x \in D} E[f(\omega, x) \mid P_P(\omega)] \right] + (1 - \rho_P) E_\omega \left[\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)] \right], \tag{8}$$

which equals $\rho_P(d + \pi) + (1 - \rho_P)\pi = \pi + \rho_P d$. □

Proof of Claim 2. Let E^{ρ_A} denote the expectations of a ρ_A-biased agent. The ex-ante expected mean estimate of π by agent A is

$$E_\omega \left[E^{\rho_A} [\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)] \mid P_A(\omega)] \right], \tag{9}$$

where $P_A(\omega)$ is the information of the particular agent A in a given state. Using the definition, this equals

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

Given the law of iterated expectations as before, this then becomes $\rho_A \pi + (1 - \rho_A)\pi = \pi$.

Consider now the expected estimate of the agent of the mean estimate of principal of π. This corresponds to

$$E_\omega \left[E^{\rho_A} [E^{\rho_P} [\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \mid P_A(\omega)] \right] \tag{10}$$

Using the definition and the linearity of expectations, we can re-write this as

$$\begin{aligned} E_\omega \Big[& \rho_A E^{\rho_A} [\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)] \mid P_A(\omega)] \\ & + (1 - \rho_A) E[E^{\rho_P} [\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \mid P_A(\omega)] \Big] \end{aligned} \tag{11}$$

The first part of the above expression is based on the fact that agent A projects onto *all* the reference agents as well and believes that the projected version of the principal has the same beliefs about the strategies of the reference agents as he does. We can re-arrange the above expression to obtain

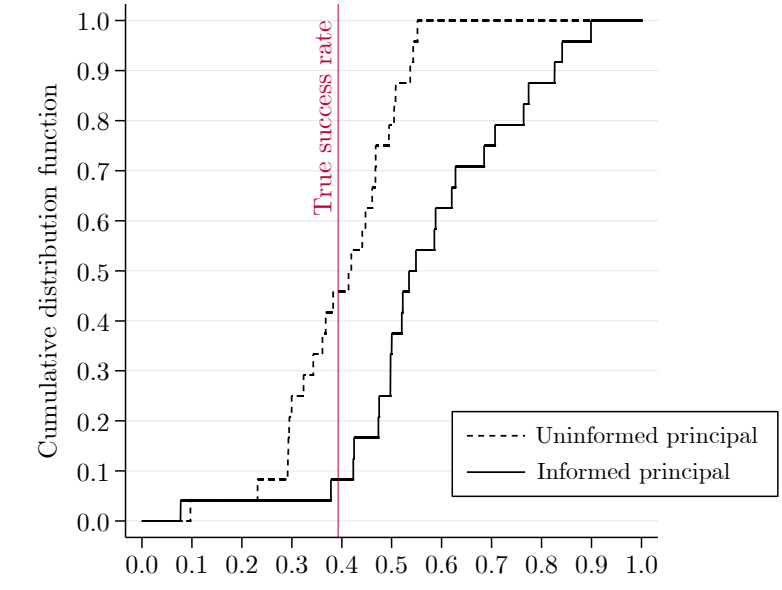
$$\rho_A \pi + (1 - \rho_A) E_\omega \left[E^{\rho_P} [\max_{x \in D} E[f(\omega, x) \mid P_{\overline{A}}(\omega)] \mid P_P(\omega)] \mid P_A(\omega)] \right]. \tag{12}$$

Given Claim 1, the second term equals $(1 - \rho_A)((\rho_P(d + \pi) + (1 - \rho_P)\pi))$. Hence the above expression equals $\pi + (1 - \rho_A)\rho_P d$ □

6.2 Supplementary analysis

6.2.1 Stated beliefs of the principals

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



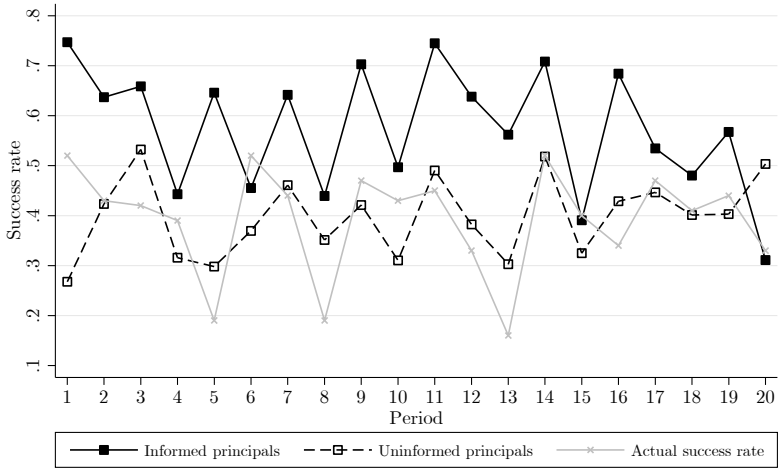
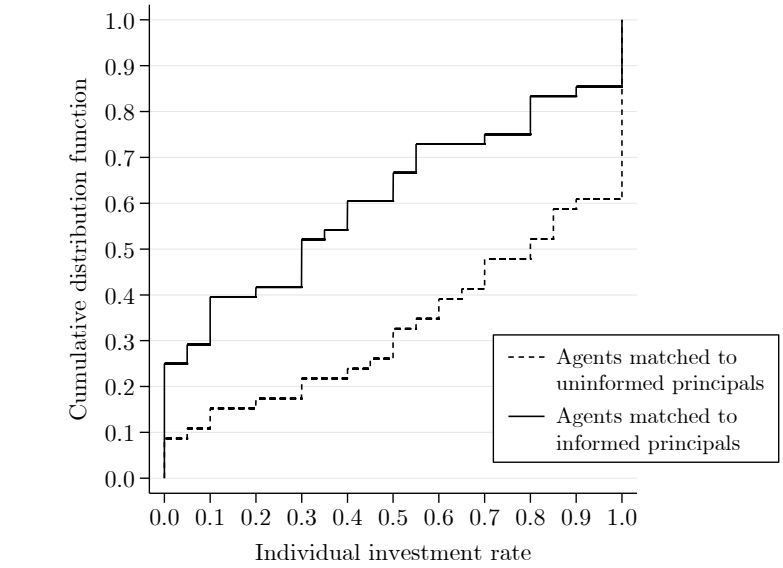


Figure 6: Average performance estimates of principals and actual success rate of the reference agents over time.

6.2.2 Investment decisions of the agents

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



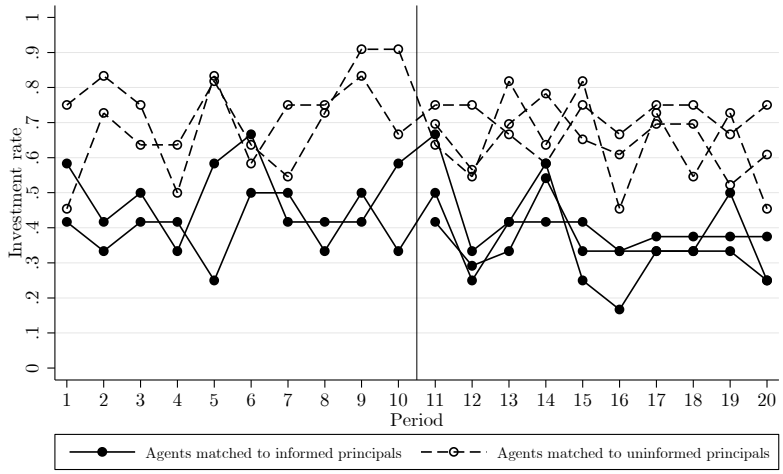


Figure 8: Investment rates per session over time.

6.2.3 Stated beliefs of the agents

Table 3: 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

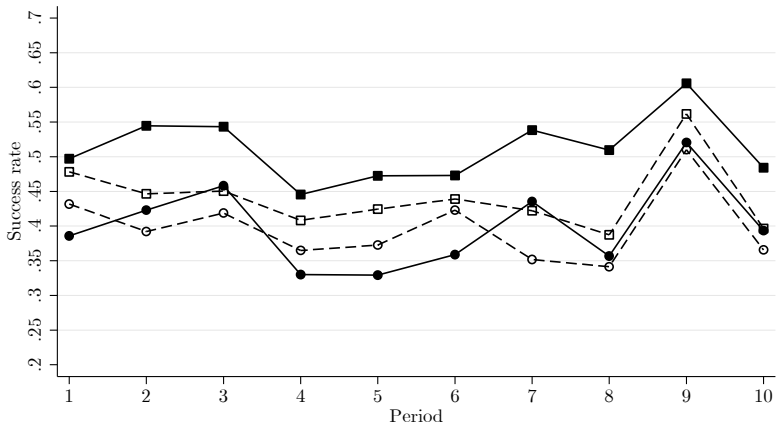
Note: Values in parentheses represent standard errors corrected for clusters on the individual level. Asterisks represent p-values: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: 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: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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.



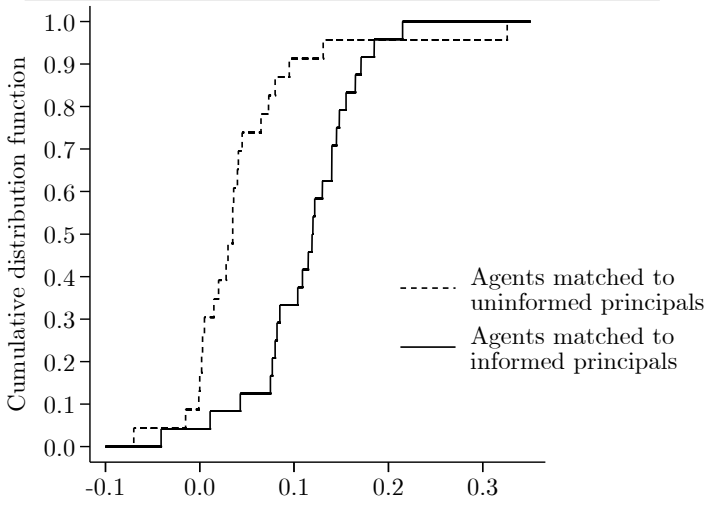


Figure 10: Empirical cumulative distribution functions of each agent's average difference between her second-order belief (estimate of the principal's estimate of the success rate) and her first-order belief (own estimate of the success rates), conditional on being matched with informed or uninformed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

Table 5: 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 6: 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 7: 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: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

6.3 Out-of-sample validation

The following section provides a simple test of projection equilibrium's out-of-sample predictive accuracy. Our design allows us to try to predict the agents' investment choices in the second half of the experiment based on their stated beliefs in the first half of the experiment. The exact procedure is as follows: for each task in the second half of the experiment, we first set the agents' first- and second-order beliefs as implied by projection equilibrium. The agents' first-order beliefs are correct, i.e., correspond to the actual success rates in each task. The predicted second-order beliefs are calculated according to (6) where we set $\rho^A = \rho^P$. We then calculate the predicted investment choice as the optimal choice given these beliefs, where the agent is predicted to invest if her predicted second-order belief does not exceed her predicted first-order belief by more than 10 percentage points.

Projection equilibrium correctly predicts the agents' investment choices 67.5% of the time which is significantly higher than the random benchmark of 53.125% (obtained by simple unconditional resampling from the empirical distribution of investment choices; $p = 0.0461$).[39] Furthermore, the predictive accuracy of the unbiased Bayesian Nash Equilibrium is only 37.5%. Note that this is the second-best model in terms of what we can expect in terms of predictive accuracy (see discussion below). The accuracy of the unbiased BNE is significantly lower than both the predictive accuracy of projection equilibrium ($p = 0.005$) and the random benchmark of 53.125% ($p = 0.0492$).

[39] The random benchmark corresponds to the expected value of randomly picking an investment choice from the empirical distribution, with replacement, and assigning it to an agent. In the second half of the experiment, informed agents (for whom we elicited beliefs) invested in 90/240 (37.5%) of the cases. The random benchmark is thus calculated as $(90/240)^2 + (1 - 90/240)^2$. The p-value corresponds to a t-test of the average individual fraction of correctly predicted choices versus the random benchmark.

6.4 Instructions

6.4.1 Instructions for principals (translated from German)

Welcome to the experiment!

The experiment that you will be participating in is part of a project funded by the German Research Foundation (DFG). It aims to analyze economic decision-making.

You are not allowed to use any electronic devices or to communicate with other participants during the experiment. You may only use programs and features that are allocated to the experiment. **Do not talk to any other participant.** Please raise your hand if you have any questions. We will then approach you and answer your question in private. Please do not under any circumstances raise your voice while asking a question. Should the question be relevant for everyone, we will repeat it aloud and answer it. If you break these rules, we may have to exclude you from the experiment and from receiving payment.

You will receive a show-up fee of 5 Euros for your attendance. You can earn additional money through the experiment. The level of your earnings depends on your decisions, the decisions of participants in former experiments, and on chance. The instructions are the same for everyone. A detailed plan of procedures and the conditions of payment will be explained below.

Tasks

You will face 20 tasks in the course of the experiment. For each task, you will be shown a short video. Each video consists of two images that are shown alternately. Your task is to spot the difference between the images.

The duration of each video is 14 seconds. After each video, you will have 40 seconds to submit an answer. The interface will show a numbered grid together with the image containing the difference. To solve the task, enter *one* of the numbers corresponding to a field containing the difference. If the difference is covered by

more than one field, each field containing the difference will be evaluated as a correct answer.

You will receive [INFORMED TREATMENT: 0.30 Euros] [UNINFORMED TREATMENT: 0.50 Euros] for each task you solve correctly.

Estimates

Other participants in previous experiments performed all tasks you will face in this experiment. Like you, these *previous performers* also had to spot the difference between the two images of each video.

After performing each task, you will have the opportunity to estimate the percentage of previous performers that spotted the difference. Therefore, you will watch exactly the same videos as the previous performers. The previous performers also had 40 seconds to submit an answer, just like you, and were also paid according to their performance. [INFORMED TREATMENT ONLY: Before each video, you will receive a guide to the solution of the task. Please note that the previous performers did not receive solution guides.]

At the end of the experiment, the computer will randomly select two videos. Your estimates for these videos will be relevant for your payoff. For each of the payoff-relevant videos, the following holds: If your estimate is within the interval +/− 5 percentage points around the true percentage of previous performers that correctly identified the difference, you receive 12 Euros.

Consider the following example. Assume that for one of the two payoff-relevant videos, 50% of the previous performers correctly identified the difference in this video. If you estimated that 53% of the previous performers spotted the difference, then you will receive 12 Euros. However, if you estimated that 57% of the previous performers spotted the difference, then you will receive 0 Euros.

You will start with three practice videos to familiarize yourself with the procedure. The practice rounds are not payoff relevant. Afterward, the 20 payoff-relevant videos will follow.

Further procedure

After you submit your estimates for the 20 videos, you can earn money with additional decision-making problems. Further details will be given during the experiment.

At the end of the experiment, we will ask you to fill in a questionnaire. Even though your answers will not affect your payoff, we kindly ask you to answer the questions carefully.

After you completed the questionnaire, you will be informed about your payoff from performing the tasks, your payoff from your estimates, your payoff from the additional decision making problems, as well as your total payoff in this experiment. Please remain seated until the experimenter lets you know that you may collect your payment.

Do you have questions? If yes, please raise your hand. We will answer your questions in private.

Thank you for participating in this experiment!

Comprehension questions

1. How many (potentially payoff-relevant) videos will you evaluate in total?

   _____

2. How many of these videos will be selected for payment regarding your estimate of the previous performers?

   _____

3. What are the components of your total payoff?

   _____

4. Assume the computer randomly selected the following exemplary videos for payment regarding your estimates

   Video a): 62% of the previous performers found the solution.

   Video b): 35% of the previous performers found the solution.

   What is your payoff from your estimates if you estimated that in video a) 57% and in video b) 36% of the previous performers spotted the difference?

   _____

5. What is your payoff from your estimates if you estimated that in video a) 87% and in video b) 29% of the previous performers spotted the difference?

   _____

6.4.2 Instructions for agents (translated from German)

Welcome to the experiment!

The experiment that you will be participating in is part of a project funded by the German Research Foundation (DFG). It aims to analyze economic decision-making.

You are not allowed to use any electronic devices or to communicate with other participants during the experiment. You may only use programs and features that are allocated to the experiment. **Do not talk to any other participant.** Please raise your hand if you have any questions. We will then approach you and answer your question in private. Please do not under any circumstances raise your voice while asking a question. Should the question be relevant for everyone, we will repeat it aloud and answer it. If you break these rules, we may have to exclude you from the experiment and from receiving payment.

You will receive a show-up fee of 8 Euros for your attendance. You can earn additional money through the experiment. The level of your earnings depends on your decisions, the decisions of participants in former experiments, and on chance. The instructions are the same for everyone. A detailed plan of procedures and the conditions of payment will be explained below.

Tasks

You will face 20 tasks in the course of the experiment. For each task, you will be shown a short video. Each video consists of *two images* that are shown *alternately*. Your task is to spot the difference between the images.

Figure 1 shows an example. Image A of the example shows a kayaker on the left side. In image B, the kayaker is not present.

The duration of each video is 14 seconds. After each video, you will have 40 seconds to submit an answer. The interface will show a numbered grid together with the image containing the difference (see Figure 2).

Image A

Image B

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

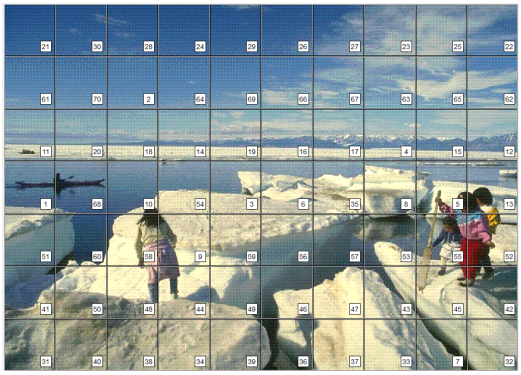


Figure 2: Response grid.

In each video, the difference between the images covers at least two fields. To solve the task, enter *one* of the grid numbers corresponding to a field containing the difference. That is, the number of any field containing the difference will be evaluated as a correct answer.

The experiment consists of 20 such tasks. The order of the tasks was randomly determined by the computer and is the same for all participants. You will receive 0.50 Euros for each task you solve correctly.

Previous participants

Performers

Participants in previous experiments performed all tasks you will face in this experiment. Like you, these *previous performers* had to spot the difference between the two images in each video. They were also paid according to their performance.

Evaluators

The degree of difficulty of the tasks that have been performed by the previous performers (and will be performed by you in this experiment) has been evaluated by further participants of previous experiments.

These *evaluators* were shown the tasks in the same way as the previous performers (and you), including the 40-seconds response time. After watching a video, the evaluators estimated the fraction of previous performers that solved this task correctly. The evaluators were paid according to the accuracy of their estimates.

[INFORMED TREATMENT ONLY: In contrast to the previous performers (and you), the evaluators received guides to the solution *before* each task. The evaluators were informed that the previous performers did not receive solution guides.]

At the beginning of the experiment, one of the evaluators will be randomly matched to you.

Insurance decision

During the experiment, you can earn additional money through insurance decisions. You will make one insurance decision after each task. At the end of the experiment, one of your insurance decisions will be randomly selected for payment.

Your endowment for each insurance decision is 10 Euros.

Not buying insurance

If you do not buy the insurance, your payoff depends on the following factors:

(1) the number of previous performers (percent) who solved the current task + *10 (percent)*,

(2) the number of previous performers (percent) who solved the current task *in the evaluator's estimation.*

If **(1)** *is at least as high as* **(2)**, you will keep your endowment of 10 Euros.

If **(1)** *is smaller than* **(2)**, you will lose your endowment; that is, you will receive 0 Euros.

In other words, if you do not buy the insurance, your payoff will be determined as follows: You will keep your endowment of 10 Euros if the evaluator's estimate of the performance of the previous performers is correct, an underestimation, or an overestimation by not more than 10 percentage points. Otherwise, you will lose your endowment.

Buying insurance

You have the opportunity to insure against this risk. The insurance costs 6 Euros. If you buy the insurance, you will receive your endowment minus the cost of insurance, that is, 4 Euros.

Example 1

Assume that 50% of the previous performers actually solved the task. In the evaluator's estimation, 60% of the previous performers solved the task. Of course, you will not learn these values during the experiment.

If you did not buy the insurance, you will keep your endowment, because 50% + 10% $\geq$ 60% The payoff from your insurance decision will therefore be 10 Euros.

If you bought the insurance, you will receive your endowment minus the cost of insurance, that is, 4 Euros.

Example 2

Assume again that 50% of the previous performers actually solved the task. In the evaluator's estimation, 20% of the previous performers solved the task. As in the previous example, you will receive 10 Euros if you did not buy the insurance and 4 Euros if you bought the insurance.

Example 3

Assume again that 50% of the previous performers actually solved the task. In the evaluator's estimation, 70% of the previous performers solved the task. In this example, you will receive 0 Euros if you did not buy the insurance and 4 Euros if, you bought the insurance.

Summary and further procedure

The experiment consists of 20 rounds in total. At the beginning of each round, you will work on the task; that is, you will receive 0.50 Euros if you spot the difference between the two images in the video.

After each task, you can earn additional money through an insurance decision. At the end of the experiment, one of your insurance decisions will be randomly selected for payment. Because you do not know which insurance decision is selected for payment, you should treat each insurance decision as if it were payoff relevant.

After you make your insurance decision, you will receive feedback with a guide regarding the solution to the current task. [INFORMED TREATMENT: The evaluators received the same guide before watching the task.] [UNINFORMED TREATMENT: The evaluators did not receive solution guides.]

You will start with five practice rounds to familiarize yourself with the procedure. The practice rounds are not payoff relevant. Afterward, the 20 payoff-relevant rounds will follow.

After the 20 rounds, you can earn money through additional decision-making problems. Further details will be given during the experiment.

At the end of the experiment, we will ask you to fill in a questionnaire. Even though your answers in this part will not affect your payoff, we kindly ask you to answer the questions carefully.

Do you have questions? If yes, please raise your hand. We will answer your questions in private.

Thank you for participating in this experiment!

Comprehension questions

1. When do you make your insurance decision—before or after you watched the task?

2. Are the following statements true or false?

   (a) The evaluators received a guide to the solution to each task.

   (b) The evaluators did not receive guides to the solution to the tasks.

   (c) The evaluators received a guide to the solution before watching and evaluating the task.

   (d) At the end of the experiment, one of my insurance decisions will be randomly selected by the computer for payment.

   (e) My payment at the end of the experiment consists of the show-up fee (8 Euros), the payoff from performing the tasks (0.50 Euros per task solved), the payoff from one insurance decision, and the payoff from further decision-making problems.

3. Assume the computer randomly selected videos for payment with the following exemplary characteristics. Provide the payoff from your insurance decision for each example.

   Example a)

   - 62% of the previous performers found the solution.
   - The evaluator randomly matched to you estimated that 72% of the previous performers spotted the difference.
   - You did not buy the insurance.

   Example b)

   - 62% of the previous performers found the solution.
   - The evaluator randomly matched to you estimated that 75% of the previous performers spotted the difference.
   - You did not buy the insurance.

Example c)

- 34% of the previous performers found the solution.
- The evaluator randomly matched to you estimated that 41% of the previous performers spotted the difference.
- You did not buy the insurance.

Example d)

- 34% of the previous performers found the solution.
- The evaluator randomly matched to you estimated that 30% of the previous performers spotted the difference.
- You did not buy the insurance.

4. Consider again the examples in 3. For each example, assume you did not buy the insurance. What is your payoff from your insurance decisions in each of the examples?