Emotional Susceptibility to Public Scrutiny and Vaccine Hesitancy: An Exploratory Experimental Analysis*

Christine Alamaa[1] ✉ and Alice Dominici[2,3]

[1]The Swedish Institute for Social Research, Stockholm University

[2]IMT School for Advanced Studies Lucca

[3]Bocconi University, Dondena Centre

1 February, 2026

Abstract

This paper explores the understudied link between self-consciousness and vaccine scepticism, combining an experimental approach with causal forests to estimate individual treatment effects. Leveraging data from a laboratory experiment with Italian university students, we find that individuals who are more easily induced to self-conscious responses (e.g., feeling shame or embarrassment) in response to public scrutiny tend to hold stronger vaccine misbeliefs. Rather than identifying a causal effect of self-consciousness elicitation on vaccine attitudes, our results highlight a correlation between pre-treatment attitudes and susceptibility to self-conscious emotions. This suggests that studying targeted public health communication may be crucial, as more sceptical individuals could avoid discussing with health professionals or develop self-conscious emotions as a result of these interactions, further exacerbating their vaccine hesitancy.

Keywords: Self-conscious emotions, vaccine hesitancy, public health policy

JEL Codes: I18, D04, D91

*We thank Erik Lindqvist, Jenny Säve-Söderbergh, Johannes Hagen, Roel van Veldhuizen, Anna Dreber, Astri Muren, Emil Persson, and Anna Sandberg for valuable comments on the manuscript. We are grateful to Natalia Montinari and Irene Buso for operational support in data collection at the BLESS Lab at the University of Bologna. All remaining errors are our own. Financial support from the Swedish Research Council (VR), the Swedish Research Council for Health, Working Life and Welfare (FORTE), and the European Union–Next Generation EU programme is gratefully acknowledged.

✉ christine.alamaa@sofi.su.se, The Swedish Institute for Social Research (SOFI), Stockholm University.

1 Introduction

Emotions such as shame, guilt, embarrassment, and pride are self-conscious responses that arise when individuals evaluate their behaviour against social expectations and the reactions of others (e.g., Tangney et al., 2007; Tracy and Robins, 2007). Among these emotions, shame has been recognised as a fundamental obstacle to successful doctor-patient interactions and healthcare provision in the medical literature (reviewed by Jaeb and Pecanac, 2024). In the economic literature, self-conscious emotions (SCEs) are viewed more broadly as non-monetary drivers of individual decision-making, particularly relevant when actions are observable to others or when individuals internalise social norms (Bénabou and Tirole, 2006; Ellingsen and Johannesson, 2008). Whenever such moral emotions, like SCEs, arise from observability, they can be powerful drivers of behaviour, motivating avoidance of social disapproval and exclusion (e.g., Andreoni and Bernheim, 2009; Fischbacher and Föllmi-Heusi, 2013), and serving as a disciplinary device (Falk and Ichino, 2006).

In this paper, we bridge these separate research strands and provide exploratory experimental evidence on the understudied relationship between self-consciousness and the policy-relevant topic of vaccine attitudes. In particular, high emphasis during COVID-19 and the high risk of future virus-borne pandemics put vaccine attitudes at the centre of heated and polarised public debate that directly informs public health policy choices, making social norms around vaccination attitudes and behaviour especially salient to policymakers relative to other health domains. In the absence of studies explicitly focused on SCEs (such as shame and embarrassment), a priori expectations are ambiguous. Self-conscious emotions could work similarly to social norms and others' expectations, which increased willingness to get vaccinated (Betsch et al., 2018), but may also lead to resistance or reactance (Barron et al., 2026), especially within a trend of decreasing trust towards scientists, doctors, and vaccines. In contexts where identity or autonomy is threatened, such as doctor-patient interactions, individuals may disengage or reject the underlying norm. This dual nature of self-consciousness, both as a compliance-inducing and as a defensive emotion, has received limited attention in the economics literature, and is likely dependent on individual attitudes.

Our analysis proceeds in two steps. First, within a larger and pre-registered lab experiment conducted on 270 Italian university students, we randomised the elicitation of self-consciousness: for treated individuals, performance in a real effort task unrelated to health, and beliefs about their

performance are made public to a principal. We find large and consistent effects on two different measures of self-consciousness, reflecting modesty and confidence bias: the former increases by 23%, and the latter decreases by 48% relative to the control group. We then use causal forests by Athey and Imbens (2016), a causal Machine Learning approach that exploits pre-treatment covariates, to recover the individual causal effect for each study participant, which indicate the individual intensity of self-consciousness. In the second exploratory phase, we present participants in the treated group with a survey module containing questions on vaccine attitudes, and conclude by asking them how much they would want to be paid to make those attitudes public to the principal. We then correlate indicators built on these answers with the individual treatment effects from the first step, uncovering the relationship between responsiveness in self-consciousness and vaccine attitudes.

We find that participants who were more responsive to public scrutiny report more sceptical attitudes towards vaccines. Our interpretation is not that the stated vaccine misbeliefs are *evoked* by self-conscious emotions such as shame or embarrassment, but rather that individuals who are more easily influenced already held more vaccine-sceptic attitudes before treatment. Supporting evidence comes from the result that, despite holding more sceptical and controversial views, these individuals are less concerned about making them public to the principal. Given the strong and consistent results of public exposure in the first phase, if the elicited self-consciousness were causally related to vaccine attitudes, we would expect stronger responses to correspond to higher vaccine self-awareness, rather than lower. This interpretation also aligns with previous research linking vaccine hesitancy to conspiracy beliefs, anti-intellectualism, and cognitive biases, suggesting that a predisposition to manipulation may play a role (e.g., Miller, 2020; Imhoff and Lamberty, 2020; Wang et al., 2025; van Prooijen and Böhm, 2024; Callaghan et al., 2019; Farhart et al., 2022; Gagliardi, 2025). Yet, the current design does not allow for disentangling the psychological mechanisms behind the observed correlation, highlighting these as hypotheses for future research in this area.

Despite relying on a small sample from a lab experiment, our findings shed light on a potentially fundamental policy challenge: individuals who are more easily induced to self-conscious emotions tend to hold stronger vaccine misbeliefs. On the one hand, this can make them less likely to seek advice from medical professionals, even when they represent the main target of public vaccination campaigns, especially in the face of potential future pandemics. Moreover, communication missteps by health professionals that inadvertently evoke self-consciousness may further alienate these

individuals, worsening their vaccine hesitancy. We find that susceptibility to self-consciousness is higher among older students, suggesting that broadening the sample to include older segments of the population may yield even stronger results.

We contribute by highlighting a possible link between self-conscious emotions and image concerns in the context of vaccine hesitancy. Given the policy challenge of promoting vaccinations, especially among the more sceptical, this uncovers an important avenue for future research in health economics and public health. While the medical literature has investigated SCEs such as shame in missed doctor-patient interactions for known unhealthy behaviours (such as smoking or drug usage, e.g., Sankar and Jones, 2005), and in relation to physical examination or illnesses (Dolezal and Lyons, 2017; Harris and Darby, 2009), shame over beliefs or attitudes has received less attention. In health economics, recent research on framed communication campaigns has focused on emotional responses to patient–provider distance, yet without explicitly considering self-conscious emotions tied to image concerns and self-evaluation. For instance, Alsan and Eichmeyer (2024) found improved flu vaccine uptake from informational videos when black, lowly educated American men with typically low vaccine acceptance were exposed to race-concordant informants. Armand et al. (2024) focused instead on the effect of religious faith concordance between informant and recipient on the uptake of preventive hygiene measures during COVID-19 in India. Closer to the topic of self-consciousness, Dominici et al. (2025) find that gentle communication techniques improve the perception of the health professional providing vaccine information without translating into higher vaccination uptake, highlighting that vaccine hesitancy is resistant to these interventions. Dominici and Dahlström (2025) show that using emotionally or scientifically-framed nudges and language can affect HPV vaccination uptake differently depending on the recipient's educational background. More generally, health economics research has acknowledged the influence of image concerns and perceived stigma on behaviours such as HIV testing, smoking cessation, and contraceptive use (e.g., Thornton, 2008; Dupas, 2011; Ashraf et al., 2014).

A complementary strand of literature on motivated beliefs further illuminates our findings. Belief distortions may arise both from motivated reasoning that protects self-image and from cognitive mechanisms such as selective attention and memory, which overweight salient information (Bordalo et al., 2020). This mechanism has been used to explain resistance to climate change mitigation, but it may also help explain why certain individuals can become more vaccine-sceptical in response to shame-based messaging. Grossman and van der Weele (2017) formalise related ideas in their

model of wilful ignorance, showing that individuals may strategically avoid information to maintain a positive self-image in morally loaded contexts. In the health domain, receiving a message that elicits self-conscious discomfort may compel individuals to reject the message altogether in order to preserve a coherent, favourable view of the self. Empirical support for this comes from Norgaard (2006), who describes socially organised denial of climate change in Norway not as a lack of awareness, but as a coping mechanism to avoid guilt and moral conflict. Similarly, Nyborg (2011) discusses how consumers deliberately avoid information about ethical consumption in order to avoid confronting the moral implications of their choices. These insights reinforce the idea that moralised messaging—such as evoking self-conscious emotions (e.g., shame or embarrassment) to promote health behaviour—can paradoxically provoke psychological resistance, especially among individuals whose identities are threatened by the implied moral judgment. Our results contribute to this literature by suggesting a novel application of these mechanisms in the context of vaccine attitudes, where emotional discomfort may lead not only to belief distortion but to deeper scepticism and disengagement.

The remainder of this paper proceeds as follows. Section 2 describes the experimental setting and our two-steps design. Section 3 introduces the data and methods used in our empirical analyses, whereas Section 4 presents and discusses the results. Finally, Section 5 concludes.

2 Context and Experimental Setting

This study draws on a health-related survey module embedded within a larger incentivised behavioural experiment. Our setting is targeted to isolate the effects of two features common to large-scale public health programs on individual health attitudes. Public health programs aimed at reducing vaccine hesitancy and promoting preventive health measures are often launched through personal interactions or targeted information programs (e.g., gentle communication techniques applied to doctor-patient interactions, now recommended by the World Health Organization— WHO, 2016). In these settings, patients' disclosure of health beliefs and habits exposes them to public scrutiny that may evoke shame, a self-conscious emotion (Jaeb and Pecanac, 2024; Dolezal and Lyons, 2017).

To examine how individual susceptibility to self-conscious emotions relates to health attitudes,

participants were randomly assigned to a control or a treatment group, both of which completed the main experimental tasks. First, we vary public exposure in the main task (Module 1) to study individuals' self-conscious responses, with no mention of any health attitudes. Second, we introduce a supplementary module designed to measure health attitudes (Module 2). This module was structured as an "add-on" to the main experimental task and, due to logistic constraints, was presented only to treated participants. In Module 1, we use causal forests to estimate individual treatment effects of public scrutiny in a novel way, exploiting random assignment between the treatment and control groups in the main task. This provides us with individual-level causal estimates (CATEs) for each treated participant, which—despite the lack of a control group in Module 2—we can then correlate with their reported health attitudes. As a result, our health-focused analysis is based exclusively on this group. This modular setup allows us to explore whether individuals who are more responsive to self-consciousness in the main task also exhibit stronger vaccine scepticism or greater discomfort in revealing health attitudes.

Self-conscious emotions. Our design elicits behavioural responses to public scrutiny through exposure, which we interpret as susceptibility to self-conscious emotions (SCEs). Self-conscious emotions are a family of emotions that arise from self-evaluation against perceived social standards, and include shame, guilt, embarrassment, and pride (Tangney et al., 2007; Tracy and Robins, 2007). Unlike basic emotions, SCEs require reflection on how one's behaviour may be seen by others, and are particularly likely to be triggered in situations involving external evaluation and social exposure. Meta-analytic evidence suggests that shame and embarrassment are robustly elicited in evaluative social contexts (Else-Quest et al., 2012). While our design does not permit measuring shame in a strict psychological sense, the structure of the task—publicly exposing one's self-assessment—makes shame or embarrassment the most plausible drivers of the observed responses.[1] In the analysis that follows, however, we adopt the broader term *self-consciousness* to describe the mechanism under study, while noting that in our setting the most plausible drivers of behaviour are shame- or embarrassment-related responses to public exposure.

This section is structured as follows: Subsection 2.1 describes the main experimental task (Module 1), where self-consciousness is elicited through public exposure and scrutiny. We then explain how we estimate individual treatment effects using causal forests. Finally, we introduce the health mod-

[1] By minimising the role of monetary stakes (through the payoff structure) and maintaining anonymity of absolute earnings, the design isolates public scrutiny as the primary varying factor.

ule (Module 2) in Subsection 2.2, and show how it is used to connect self-conscious responsiveness to health-related attitudes. Further methodological details are provided in Section 3.

2.1 Experimental Design: Public Scrutiny in Module 1 (full sample)

This module is designed to elicit self-consciousness by exposing the (in)accuracy of participants' self-assessments to an external observer. We vary the degree of observability across treatments to study behavioural responses to public scrutiny, which we interpret as reflecting susceptibility to self-conscious emotions. The psychological mechanism underlying this design is rooted in the social nature of self-conscious emotions: individuals may experience emotional discomfort when they believe others can observe a discrepancy between their self-evaluation and objective performance. In particular, overestimating one's relative standing—especially when this confidence bias becomes visible to others—can evoke feelings of shame or embarrassment, which are SCEs (Tangney and Fischer, 1995; Tesser, 1988). This tendency is amplified when self-assessment conflicts with norms of modesty or accuracy, and when individuals are subject to public scrutiny. By making the inaccuracy of rank beliefs observable to a passive authority figure (the principal), we create conditions under which such self-conscious responses are likely to arise.

The experiment involves two participant roles: principals and agents. Each principal is paired with three agents, forming a group of four participants. Roles and group assignments were randomised and remained fixed throughout the experiment.[2] Each session included 24 participants—18 assigned to the agent role and 6 to the principal role. Principals play mainly a passive role, observing the decisions of the three agents assigned to them. Our analysis focuses solely on agents (hereafter referred to as participants). Module 1 has a repeated structure, collecting data on agents' choices across three identical rounds. We implement two conditions—a control and a treatment—that vary the observability of participants' self-assessment accuracy to their principal.[3] The three identical rounds unfold over four stages:

Stage 1: A 4-minute Real-Effort Task. Agents work independently on a real-effort task (RET), solving as many problems as possible within four minutes. The problems are unrelated to health

[2] In the larger experiment of which this study is a part, we adopt a standard labour-market framing: agents are referred to as "Employees" and principals as "Principal."

[3] This variation serves as the foundation for estimating individual-level self-consciousness effects, later linked to vaccine attitudes.

attitudes and involve decoding a 5-digit key into a matching 5-letter string (*see* Figure A.1 in Section A in the Appendix for an example). Each correctly solved problem adds one point to the total task score, which constitutes one of the two components used to determine participants' earnings.[4]

Following the real-effort task, agents' scores within each experimental session are ranked, but the rank is not revealed to the subjects.[5] In the next stage, we elicit agents' rank beliefs by having them invest Experimental Currency Units (ECUs) in possible session ranks, r_i, from the list $R \equiv \{r_1, r_2, \ldots, r_{18}\}$, in two steps.

Stage 2: An Incentivised Rank-Belief Elicitation Task. Agents are endowed with 19 ECUs, which are non-storable and non-transferable. They can invest their 19 ECUs in a single rank or distribute them across several ranks, provided that all units are used.

> *Step A:* Agents first choose which rank(s) r_i to invest in under the condition that they select at least one, and at most eighteen ranks, so that $i \in [1, 18]$.
>
> *Step B:* Agents then decide how much to invest in the ranks chosen in Step A, with the requirement that one rank receives at least one ECU more than any other. This ensures that one rank is clearly preferred among the selected options. Formally, if $a(r_j)$ denoted the ECU allocation to rank r_j, and $\hat{r}^*$ the most preferred rank, then the condition requires that $a(\hat{r}^*) \geq a(r_j) + 1$, for all $r_j \neq \hat{r}^*$.[6,7]

Earnings. The rank-belief elicitation task is incentivised by design: it determines an agent's bonus pay jointly on their exerted effort in the real-effort task (RET) and the accuracy of their rank belief. Rather than paying out separately, the two components are contingent on each other through multiplication. Formally, if an agent allocates $a(r_i)$ ECUs to rank r_i, this amount is multiplied by their RET score: $\textit{round earnings} = a(r^*) \times \textit{score}_{RET}$, where r^* is the agent's true rank. Although a selected rank yields a non-zero multiplier only if it matches the agent's true rank, earnings can come from any of the selected ranks, not only the most preferred one. If no ECUs are placed on

[4] After the real-effort task, subjects state their score beliefs—knowing the number of problems they attempted—before their score is privately revealed through an incentivised score-guessing task.

[5] We assigned ranks using a dense ranking rule: after a tie, the next item is given the immediately following rank number. Equivalently, each item's rank is 1 plus the number of distinct higher scores.

[6] Agents can revise their allocations in Step B and return to Step A if preferred.

[7] We define $\hat{r}^*$ as the agent's most preferred rank belief (the rank to which they allocated the most ECUs) and r^* as the agent's true performance rank, based on their relative score in the session. The difference $\hat{r}^* - r^*$ captures both the direction and magnitude of belief accuracy.

the correct rank, or if the RET score is zero, the agent receives no bonus.

The principal's payoff equals the average earnings of their three assigned agents, thereby aligning material incentives between agents and principals. This payoff linkage is identical in both the control and treatment conditions.

Stage 3: Receiving Private Feedback. Agents privately receive feedback that summarises their real-effort score and their rank-belief allocations in a table format, but not their true rank. The table also highlights the rank to which they allocated the most ECUs.

Stage 4: Observability of Personal Information. In the final stage, we implement the information treatment. Agents send information to their principal using a submission form that varies by treatment condition.

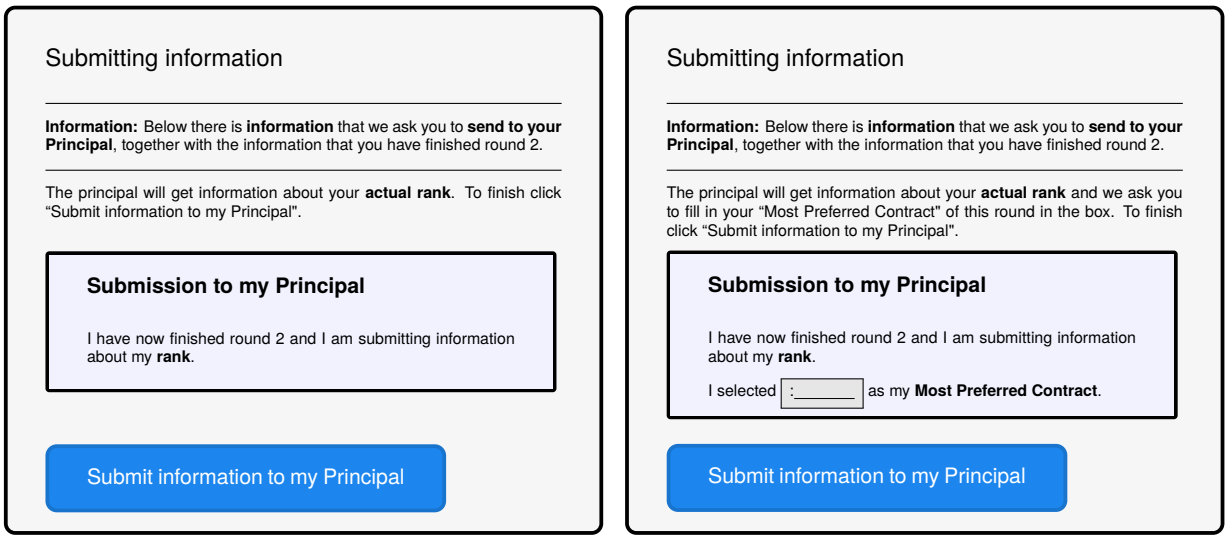
In essence, the submission forms differ in whether they reveal the accuracy of agents' rank assessments, as expressed through their choices in Stage 2:

> ***Control****:* Agents send information on their true rank, r^* (which is unobserved to them), from that round. Principals, therefore, see the true ranks of all three of their agents in a table format.
>
> ***Public treatment****:* Agents submit information on the *true rank*, generated by comparing the real-effort score amongst agents within a session (just as in the *Control*). In contrast to the *Control*, agents also report the unique rank in which they invested the most ECUs—their modal rank belief $\hat{r}^*$. Principals therefore receive a summary table showing each agent's true rank, their reported (modal) belief about their rank, the numerical difference between the two (e.g., –2), and a labelled interpretation ("Underestimation," "Overestimation," or "Accurate estimation").[8]

Figure 1 displays how agents disseminate information on rank outcomes and modal rank-beliefs to their principals in *Stage 4*. In the left panel, Figure 1**a.** is the *Control* in which principals cannot assess the accuracy in agent-beliefs. By contrast, the right Figure 1**b.** shows agents' submission form from the *Public treatment* condition, where they additionally provide their modal rank-beliefs.

[8] We compute the difference as the true rank number minus the modal rank.

Figure 1: Example of experimental instruction in round 2—by treatment



(a.) Control **(b.)** Public Treatment

NOTES: The left panel Figure 1**a.** displays the submission-form used for the *Control* condition, in which they submit to the principal their true rank relative to other agents (unobserved to them). The right panel Figure 1**b.** shows the submission form used for the *Public* treatment, where on top of the unobserved true rank, agents submit their own incentivized self-assessment, i.e., a guess of their rank. Agents are only paid if their guessed rank corresponds to their true rank (on top of an unconditional show-up fee).

Importantly, we emphasise that agents in neither the *Control* nor the *Public treatment* condition ever learn their true rank. They are asked to submit rank-related information without knowing how their performance compares to the performance of others, and they never observe the real-effort scores of fellow participants.

Incentive Alignment and Belief Elicitation. As a final note on the experimental design, we clarify two key design choices. First, our belief-elicitation method allows participants considerable flexibility to express their beliefs by allocating their endowment across the entire range of ranks to reflect their confidence distribution.[9] We impose that one rank be expressed most confidently, receiving at least 1 out of 19 ECUs more than any other. While we acknowledge that this constraint may not perfectly capture the beliefs of a completely uncertain participant, it is essential to the public-exposure mechanism. It produces a unique most-preferred rank that can be compared to the actual rank and, in the treatment condition, shown to the principal, while still allowing earnings from any

[9]This is subject only to the rank space being discrete and a marginal unimodality requirement (agents distribute 19 ECUs across 18 ranks).

other positively allocated rank. Second, the incentive structure embedded in this design is twofold: it is non-competitive among agents and aligns incentives across roles. That is, all agents can, in principle, earn simultaneously based on their own performance and rank-belief selection, and when agents maximise their own expected earnings by allocating ECUs in line with their belief about the true rank, they also maximise their principal's earnings. At baseline, principals observe only each agent's true rank and, at the end of the experiment, their total earnings from all three agents. In the treatment condition, principals additionally see each agent's most preferred rank, which enables them to assess the accuracy of the self-assessment. However, payoffs to both principal and agent can still come from any rank to which the agent allocated ECUs. We acknowledge that, in theory, an agent adjusting their allocation to reduce the risk of public exposure as overconfident may also consider the negative externality this imposes on their principal's earnings, in addition to lowering their own potential income. Such "other-regarding" concerns would work in the opposite direction of the direct self-consciousness effect, dampening its net impact. However, since earnings can only come from a rank equivalent to the true rank, the risk is low that agents substantially reallocate to ranks they believe to be very unlikely. At most, they may adjust the relative amount placed within the same set of plausible ranks, in effect changing the exposed most preferred rank.[10] Any such countervailing effect would dampen the direct self-consciousness impact, making our estimates a lower bound. Moreover, the high degree of anonymity in the interaction, combined with the fact that income can be generated from any selected rank, makes it unlikely that this mechanism materially influences our results.

Having described how we elicit self-consciousness through public exposure of self-assessed relative performance, we now turn to the health module, where we examine how these responses relate to vaccine beliefs and willingness to reveal them publicly.

2.2 Non-Experimental Module 2: Health Attitudes and Secrecy (reduced sample)

Following the main task in Module 1, participants in the *Public treatment* completed an additional exploratory survey module on vaccine attitudes (Module 2). The module investigates how individuals with varying self-conscious responsiveness—elicited through public scrutiny in Module

[10] In our main measures of susceptibility to self-consciousness, such as "modesty" and "self-confidence," we focus on differences in rank steps—comparing the most-preferred rank to the actual rank—rather than on the amount or share allocated to a rank. See Subsection 3.2 for details.

1—express vaccine-related beliefs, and how willing they are to expose those beliefs to others. Our study is embedded in a larger experimental project in which control and other treatment arms were randomised. For logistical and timing reasons, Module 2 could only be implemented in a later data-collection wave consisting only of subjects in the *Public treatment* arm. This wave was conducted in separate experimental sessions that did not include control subjects. As a result, answers to Module 2 are not incentivised. Despite these constraints, the module provides valuable insight into how private attitudes may be shaped by concerns about social exposure. This approach is consistent with recent work discussing that incentivised belief elicitation mechanisms might not always outperform unincentivised methods in capturing truthful responses (Danz et al., 2022; Schotter and Trevino, 2014; Trautmann and van de Kuilen, 2015).

Participants were asked to respond to three statements presented simultaneously, each reflecting a distinct dimension of vaccine scepticism.[11] These statements—referred to as Statements 1–3 below—correspond to widely documented misbeliefs about vaccines, including concerns about immune system weakening, disease causation, and severe adverse effects. Similar items have been used in prior studies to measure vaccine hesitancy across cultural contexts (Betsch et al., 2018; Kata, 2010; Salali and Uysal, 2020; Dominici and Dahlström, 2025). These responses allow us to measure heterogeneity in health-related beliefs and explore their correlation with self-conscious responsiveness to public scrutiny. Responses were given on an 11-point scale from 0 ("Strongly disagree") to 10 ("Strongly agree"). We refer to these items as *Statements 1–3* in the analysis that follows:

> *Statement 1 (S1):* "Vaccines weaken the immune system."
>
> *Statement 2 (S2):* "Vaccines can cause the illness they are meant to prevent (e.g., the flu shot gives you the flu)."
>
> *Statement 3 (S3):* "Vaccines could cause serious and permanent side-effects."

After the three statements were answered, we computed the sum of the responses (ranging from 0 to 30), which was shown privately to each participant on their screen. To complement the belief items, participants indicated how many ECUs they would require to make their three responses publicly visible to their matched principal. Despite the absence of monetary incentives, this question serves as a proxy for preferences about disclosing personal and potentially sensitive health attitudes. In this

[11] A participant could not revise previous responses once recorded.

exploratory module, the reported *willingness-to-conceal* (WTC) captures how strongly individuals seek to protect their views from others.

> *Question 4 (WTC):* "How many ECUs in compensation would you accept in exchange for making your answers to questions 1, 2, and 3 above public to your principal?
>
> Recall that 1 EUR corresponds to 10 ECUs."

We interpret this "willingness-to-pay for privacy" as a measure of anticipated discomfort or self-consciousness associated with social disclosure. In real-world terms, it parallels the reluctance some individuals may feel when asked to discuss sensitive vaccine attitudes in public settings or with healthcare professionals. Participants entered a value between 0 and 500 ECUs (equivalent to €0–50), where 10 ECUs correspond to €1.

2.3 Implementation and Procedures

The hypotheses were tested using a fully computerised experiment conducted at the Bologna Laboratory for Experiments in Social Science (BLESS) during 2023 and 2024. The experiment was implemented in oTree© (Chen et al., 2016), and subjects were recruited through ORSEE (Greiner, 2015), an online recruitment platform for experimental participants.[12] Prior to participation, all subjects provided informed consent, covering data use, storage, and anonymity provisions. A total of 270 participants were assigned to the agent role in either the *Control* condition (144 participants) or the *Public treatment* condition (126 participants).[13,14] While Module 1 was pre-registered on the Open Science Framework as part of a larger project (Alamaa, 2023, June 11), Module 2 (on health attitudes) was exploratory, i.e., not pre-registered. The full experimental procedure took approximately 50 minutes to complete. An additional 20 minutes were allocated at the beginning of each

[12] At the time of data collection, the ORSEE subject pool at the University of Bologna included approximately 6,700 registered student participants.

[13] The experiment is part of a larger experimental program comprising 768 participants in total. In that broader structure, the *Public treatment* is primarily used to examine gender differences in sensitivity to public exposure. The health attitude module, however, is exclusively analysed in the present study.

[14] Randomisation was conducted at the individual level within sessions of 24 participants (18 agents and 6 principals). The allocation aimed to assign either 3 or 6 agents (plus 1 or 2 principals) to each treatment arm, based on combinatorial feasibility within sessions. Due to a coding error in the information conditions, data for the *Public* treatment had to be recollected across eight additional sessions.

session for reading instructions and familiarising participants with the interface. During this phase, subjects completed a 2-minute trial of the real-effort task, a 4-minute trial of the rank-selection task, and were required to pass a five-question comprehension check on the belief elicitation mechanism. The experiment was conducted in Italian. An English translation of the instructions is provided in Section F of the Appendix. Participants received a flat participation fee of either €5 or €10.[15] This variation was unrelated to our study design and arose from scheduling requirements within the larger experiment of which this study is a part (e.g., to achieve gender-balanced samples and to exclude any retakers); it was effectively random across sessions and applied equally to treatment and control groups. Bonus payments were determined by randomly selecting one of the three rounds in Module 1 for payoff, with average earnings of 27.96 ECUs (equivalent to €2.8). While the average realised bonus was smaller than the fixed show-up fee, the incentives participants faced were considerably larger. A simple illustrative calculation based on typical performance of 16 points gives expected earnings ranging from €1.6–€3.2 for highly risk-averse allocations (selecting all 18 ranks) to around €6.4 for moderately diversified choices (selecting about 5 ranks), and up to €30.4 for a fully risk-neutral strategy allocating all 19 ECUs to one rank.

3 Data, Methods and Conceptual Interpretation

This section describes the dataset and empirical approach used to estimate the effects of public exposure and self-consciousness on self-assessed rank and health-related attitudes. We begin by detailing the sample and variables used in the empirical analyses, then present our estimation of average and conditional treatment effects using OLS regressions and causal forests, respectively. We close the section with a simple conceptual model linking the two modules, to simplify the interpretation of the results presented in Section 4.

Our empirical analysis proceeds in two steps. First, we estimate the average treatment effect (ATE) of public scrutiny on two indicators of self-consciousness based on participants' self-assessed performance using data from Module 1. Second, we use causal forests to estimate individual-level conditional average treatment effects (CATEs), which we then correlate with health attitudes collected in Module 2.

[15] Compensation levels followed the laboratory's standard practices for session duration and were not guided by the authors.

3.1 Sample and Empirical Measures

The dataset includes 262 participants from 39 sessions conducted at BLESS.[16] Each subject completed three identical experimental rounds in Module 1. The *Control* group (141 participants) was recruited across thirty-two sessions, and the *Public* treatment group (121 participants) across seven additional sessions.[17]

To ensure consistency in the analysis, we exclude all observations from the first round. This decision is motivated by three considerations. First, the timing and information structure in Round 1 differ slightly from subsequent rounds. Because it was participants' first exposure to the experiment after reading the instructions and trying the decoding task, their responses in Round 1 may partly reflect adjustment or learning rather than stable behaviour. Second, Rounds 2 and 3 instead reflect experience-based responses, after participants had already publicly exposed their choices in Round 1 to their principal. Third, the health-attitude questions in Module 2 were asked only after all three rounds had been completed, making the later behavioural responses more salient to those answers. Importantly, all participants completed all three rounds, implying the absence of attrition. Together, these considerations make Rounds 2 and 3 the most appropriate basis for our analysis.[18]

3.2 Estimating Average Treatment Effects (ATE)

We begin our empirical analysis by estimating the average treatment effect (ATE) of public scrutiny on two individual-level outcomes constructed from Module 1:

1. **Modal Rank Belief (mrb), "modesty" [scale: 1–18]:** The rank to which the participant allocated the most ECUs in the incentivised belief elicitation task. Although participants could spread ECUs across multiple ranks, they were required to assign strictly more ECUs to one rank than any other, thereby revealing a unique modal belief. We refer to this as the participant's modal rank belief (mrb). In figures and throughout the paper, we interpret this

[16] BLESS refers to the Bologna Laboratory for Experiments in Social Science at the University of Bologna.

[17] We exclude 8 participants who strategically hedged their income by distorting effort in both Rounds 2 and 3. These distortions could compromise the incentive compatibility of the belief-elicitation mechanism (Subsection 2.1). All main results remain robust when these participants are included, though the estimates are slightly noisier. Full results with the complete sample are available upon request.

[18] Results remain quantitatively comparable, albeit slightly more imprecise, if we do not exclude the first round. While we do not present estimates from this alternative analysis, its results are available upon request.

as a behavioural measure of *modesty*. Only the rank matching the participant's actual rank yielded a payoff.[19]

An mrb (modesty) value of 1 indicates that the participant believed they ranked first in their session; a value of 18 implies they believed they ranked last. Lower values of mrb reflect higher self-assessed relative performance.

2. **Belief Error (placement), "confidence bias" [scale: −17 to 17]:** This variable measures the difference between an agent's true rank and their modal rank belief (mrb). Positive (higher) values indicate overplacement: the participant believed they ranked better than they actually did, while negative (lower) values indicate underplacement. We interpret and refer to this variable as a behavioural measure of *confidence bias*.

For each participant, we average the modal rank belief (mrb) and belief error across Rounds 2 and 3. This provides one observation per individual and reduces noise from within-subject variation. To estimate average treatment effects, we run the following ordinary least squares (OLS) regression:

$$Y_i = \alpha + \tau \cdot \text{Treated}_i + \beta \boldsymbol{X}_i + \varepsilon_i$$

where Y_i is either the modal rank belief (modesty) or belief error (confidence bias), Treated_i is a binary indicator equal to 1 if the participant was assigned to the *Public treatment*, and $\boldsymbol{X}_i$ is a vector of pre-treatment covariates from the survey. The covariate vector $\boldsymbol{X}_i$ includes the following pre-treatment characteristics:

Female: A gender indicator equal to 1 if the participant is female and 0 if male.

Age: The participant's age in years.

Higher education: The number of completed years of academic studies in higher education.

Languages spoken: The total number of languages the participant reports being able to speak, including Italian.

Risk-tolerance: A binary indicator equal to 1 if the participant chose the riskier option in a post-treatment lottery.

[19] In the experiment, participants selected among "contracts", each corresponding to a specific performance rank. We refer to these as "ranks" throughout the paper for clarity.

Learning: The change in real-effort score between Round 1 and Round 2, capturing within-task learning or adjustment.

Field of studies: This category includes three binary indicators for the participant's academic study-field major:

STEM major: Indicator for majoring in science, technology, engineering, or mathematics.

Economics/Statistics major: Indicator for majoring in economics or statistics.

Health Sciences major: Indicator for majoring in medical or health-related fields.

Covariate balance between treatment and control groups is shown in Appendix Section B, and summary statistics are reported in Figure C.1 (Appendix Section C).

3.3 Estimating Individual Treatment Effects via Causal Forests

To estimate treatment effects at the individual level, we use causal forests—a supervised machine learning method developed by Athey and Imbens (2016) and implemented as in Athey and Wager (2019). Causal forests estimate Conditional Average Treatment Effects (CATEs), allowing us to examine how the effect of public scrutiny varies with baseline participant characteristics.

In our setting, all participants in the Public treatment group were exposed to public scrutiny. However, the behavioural response to that exposure may differ across individuals. We use causal forests to estimate the individual-level treatment effect for each participant, using outcomes from Module 1 and the nine pre-treatment covariates described above (sample size: 244). Because Module 2 on vaccine attitudes was administered only to treated individuals, we cannot compare treated and control outcomes directly. Instead, we correlate the CATEs obtained from Module 1 with health attitudes from Module 2. This allows us to assess whether individuals who were more responsive to public scrutiny—through stronger self-conscious responses—are also more sceptical of vaccines or more reluctant to share vaccine-related views.

We estimate CATEs using the "honest" causal forest approach. The data is randomly split into two subsamples: an estimation sample and an inference sample. A forest of decision trees is grown on the estimation sample, where splits are chosen to maximise heterogeneity in estimated treatment effects

across leaves. For each observation in the inference sample, the CATE is computed as the average treatment effect among the other individuals in the same leaf. This ensures that no observation is used both to form the tree and to estimate its own treatment effect, reducing overfitting and bias. Section D of the Appendix provides a more formal description of the method and reports effective sample sizes and the share of treated units used to estimate each CATE. The results show that both our causal forests are characterised by good overlap and the absence of small leaves.

3.4 Linking Self-Consciousness to Health Attitudes

To explore whether self-conscious responsiveness to public scrutiny is associated with vaccine attitudes, we correlate individual-level treatment effects estimated via causal forests (CATEs) with responses from the Module 2 health questionnaire. This analysis is limited to treated participants, as Module 2 was administered only to those in the *Public treatment* condition. Since CATEs already account for observed heterogeneity through the covariates used in the forest, we run bivariate OLS regressions of the following health-related outcomes on CATEs.

1. **Misbelief: illness causation, [scale: 0, 1]:** Equals 1 if the participant agreed with Statement 2, that vaccines can cause the illness they are intended to prevent. We code agreement equal to 1 if the subject selects a value strictly greater than the midpoint (i.e., > 5, on an 11-item Likert scale).

2. **Misbelief: adverse effects, [scale: 0, 1]:** Equals 1 if the participant agreed with Statement 3, that vaccines can cause serious and permanent side effects. We code agreement equal to 1 if the subject selects a value strictly greater than the midpoint (i.e., > 5, on an 11-item Likert scale).

3. **Aggregate misbeliefs: sum score, [scale: 0–30]:** The total sum score of agreement with the Statements 1, 2, and 3, each measured on a 0–10 Likert scale, yielding a total ranging from 0 to 30.

4. **High health-secrecy [scale: 0, 1]:** Equals 1 if the participant's requested compensation to make their vaccine attitudes public (Question 4) falls in the top quartile. We refer to this measure as health-secrecy—a behavioural proxy for participants' desire to conceal their vaccine-related attitudes.

Together, these outcome measures capture both vaccine-related misbeliefs and participants' willingness to conceal their views, corresponding to the belief and vaccine-secrecy items introduced in Section 2.2. We exclude Statement 1 ("Vaccines weaken the immune system") as an independent outcome because of low response variance.[20] To minimise the influence of skew and outliers in a relatively small sample ($N = 121$), we use binary indicators for agreement rather than the full 11-point scale. In fact, responses to each belief statement are concentrated on one or two values, while rare values risk exerting disproportionate leverage, which could bias estimates. In our main analysis, we define agreement as choosing values strictly above the midpoint (i.e., > 5), where the midpoint was clearly visible to participants (Figure F.34 in the Appendix shows the screen visual). As a robustness check, we repeat the analysis on single misbelief scales excluding the value of 6 (i.e., agreement defined as > 6), and find qualitatively and quantitatively comparable results (reported in Table G.4 in the Appendix). Similarly, these considerations negatively affect the reliability of the misbelief sum score; first, the three statements are included not as binary agreements but across the full Likert scale, and the high concentration on a few values is likely to reduce the variance of the sum score, hence its effectiveness in detecting differences in beliefs; second, the sum score assigns equal weight to the three statements, even though Statement 1 suffers from extremely low variability—as discussed—and, in theory, the association between CATEs and the sum score may be primarily driven by some specific statements. We nevertheless report the sum score for the sake of completeness, alongside the results for Statements 1 and 2. Figure 2 illustrates the distribution of willingness to pay for vaccine-secrecy, and Table 1 reports summary statistics for all vaccine-related outcomes. Full response distributions are shown in Appendix Section C.

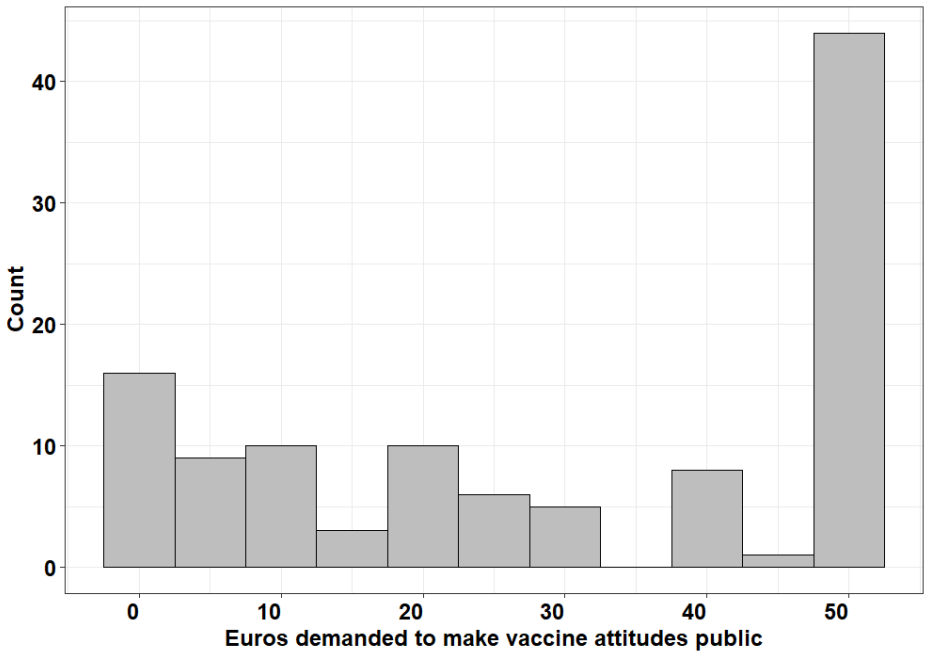
Table 1: Vaccine attitudes outcomes: summary statistics

	Likert-score > 5	
Statement	**% who agreed**	**N who agreed**
(S1): "Vaccines weaken the immune system"	5%	6
(S2): "Vaccines can cause the illness they are meant to prevent"	17%	19
(S3): "Vaccines could cause serious and permanent side-effects"	26%	29

NOTES: Agreement with each statement is measured on a 0–11 Likert scale, where 5 is visualised as the midpoint, indicating indifference. We code agreement in case of answers strictly above 5.

[20] In the case of Statement 1, only six participants agreed, making it impossible to estimate any meaningful regression.

Figure 2: Full distribution of health-secrecy measure



NOTES: The figure shows the distribution of health-secrecy, measured as the amount of monetary compensation participants would require to make their vaccine attitudes—defined as the aggregate sum score of misbeliefs (Statements S1, S2, and S3)—public to their principal.

3.5 A simple conceptual framework

We present a stylised one-shot model to contextualise the empirical results in Module 1. The agent's overall utility combines two elements: a monetary payoff linked to accuracy in beliefs, and a social cost that may arise if miscalibration becomes publicly visible. Each agent has a true rank r^* and reports a modal rank $\hat{r}$, derived from a subjective belief allocation $I(r)$ over all possible ranks. Monetary payoffs are proportional to the belief mass placed on the true rank:

$$U^m(r^*) \;=\; \kappa \cdot I(r^*),$$

where $\kappa > 0$ is a scale factor. Thus, even if $\hat{r}$ is incorrect, the agent earns as long as $I(r^*) > 0$. When belief accuracy is *publicly* revealed, visible miscalibration can generate a perceived cost stemming from an altered social-image (a reputational cost). We capture this with a single penalty that is active only under public exposure:

$$U^{\text{social}}(\hat{r}, r^*) \;=\; -\Big[\gamma^+ (r^* - \hat{r})^2 \cdot \mathbb{1}\{r^* > \hat{r}\} + \gamma^- (\hat{r} - r^*)^2 \cdot \mathbb{1}\{\hat{r} > r^*\}\Big] \cdot \mathbb{1}\{\text{public exposure}\}.$$

where $\gamma^{\pm} \geq 0$ are "mismatch sensitivities" (susceptibility to self-conscious emotions) for overplacement versus underplacement.[21] The indicator functions ensure that a reported modal rank cannot be both overplacing and underplacing; only one of the two costs can apply in any given case. The agent chooses $I(r)$ (and thus $\hat{r} = \arg\max_r I(r)$) to maximise

$$U(\hat{r}, r^*) \;=\; U^m(r^*) + U^{\text{social}}(\hat{r}, r^*).$$

Interpretation and implications. In the control condition (with no public exposure), the social term is turned off, and reporting follows monetary incentives; while under public exposure, the perceived social penalty is active. Under public exposure, the social penalty is active. If *overplacement* is reputationally the most costly ($\gamma^+ > 0$), agents optimally "shade" their modal report towards safer (more conservative) ranks or diversify belief mass away from overconfident peaks; if instead *underplacement* is perceived the most costly ($\gamma^- > 0$), the shading reverses. Hence, higher susceptibility to self-conscious emotions (larger γ^+) implies a higher (worse) reported $\hat{r}$ in the public condition.

Two points follow for the empirics. First, the observed change in reported rank under public exposure is a *net behavioural adjustment* to self-consciousness (and may be dampened by other factors, e.g., other-regarding concerns), so it should be read as a conservative lower bound. Second, heterogeneity in $\gamma^{\pm}$ across individuals naturally maps into heterogeneous treatment effects (CATEs): more susceptible agents exhibit larger downward adjustments of overconfident reporting under public exposure. These implications motivate our focus on (i) modesty and confidence bias in Module 1 and (ii) correlations between responsiveness in Module 1 and health attitudes in Module 2.[22] A possible extension of the framework is to incorporate how susceptibility to self-consciousness affects untruthful communication (e.g., by introducing an additional parameter for misreporting incentives), but such an extension is beyond the scope of this paper, as we do not have a concrete vaccine scenario or target population in mind. The latter are necessary to inform cost-benefit and welfare analyses that would derive from such an extension. In what follows, we turn to the empirical results, showing how the effects of public scrutiny on self-consciousness manifest in average treatment effects (ATEs) and heterogeneous individual responses (CATEs).

[21] Formally, the utility expression is written in net terms. While an agent may fear both being seen as overconfident and underconfident, in practice only the dominant perceived cost applies to the chosen modal rank, which determines the behavioural adjustment.

[22] A more detailed theoretical disposition, with belief updating, private feedback, and additional heterogeneity analyses, is provided in Chapter 1 of Alamaa (2025).

4 Results and discussion

4.1 Module 1: Public Scrutiny and the Elicitation of Self-Consciousness

Table 2 shows that the *Public treatment* effectively elicits self-consciousness in the full sample ($N = 244$), according to both outcome measures. First, public exposure to the principal increases agents' modesty: participants lower their self-assessed rank (modal rank belief) by more than one position. Second, overestimation of relative standing (confidence bias) decreases by more than one position. When controlling for pre-treatment covariates (Columns 2 and 4), these changes correspond to a 23% increase in modesty and a 48% decrease in confidence bias relative to the control group. Next, we compute individual-level conditional average treatment effects (CATEs) for both outcome measures: modal rank belief (modesty) and placement (confidence bias). The distributions of the two CATEs, shown in Figure 3, are consistent with the average treatment effects estimated by OLS and underline the strong statistical significance of the public scrutiny effect. Indeed, the distribution for modesty is entirely positive, and the distribution for confidence bias is entirely negative. Both exhibit substantial variation across individuals, providing statistical power for the correlations with vaccine-related outcomes estimated in the next section.

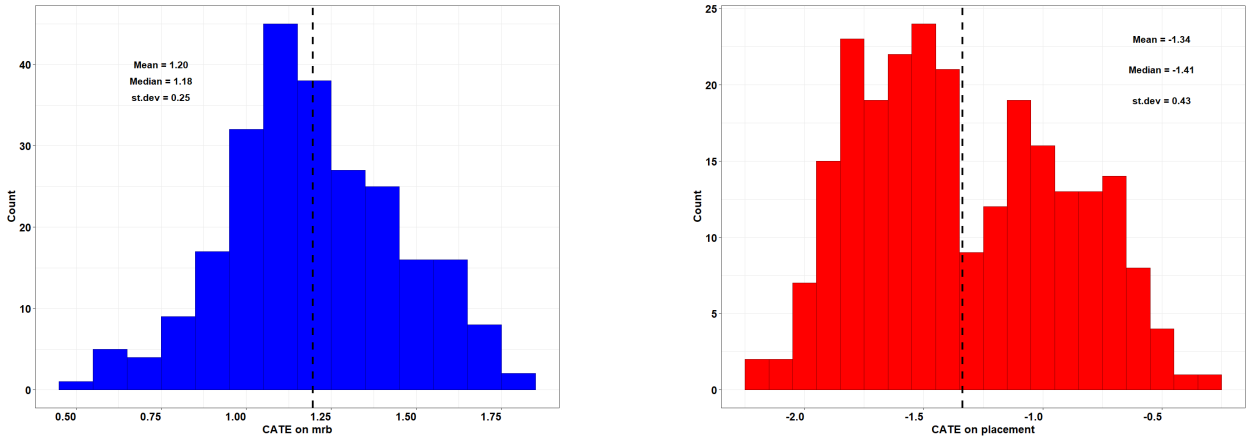
For confidence bias, the distribution shows two notable concentration peaks, with the most frequent values located beyond the average treatment effect. We also find a strong negative correlation between the two CATEs. As shown in Figure 4, the Pearson correlation coefficient is -0.603 and statistically significant at the 1% level, confirming that participants who become more modest also tend to become less overconfident in their self-assessments. Figure E.10 in the Appendix also shows that the distributions of both CATEs are comparable across *Control* and *Public treatment* participants, consistent with successful randomisation, and implying that those exposed to vaccine questions in Module 2 are representative of the full sample.

Table 2: Average Treatment Effects (ATE) of shame elicitation

	mrb "modesty"		**placement** "confidence bias"	
	(1)	(2)	(3)	(4)
Public treatment	1.515***	1.210***	-1.580**	-1.442**
	(0.534)	(0.516)	(0.686)	(0.701)
N	244	244	244	244
Covariates	-	✓	-	✓
Control mean	5.22	5.22	3.03	3.03

NOTES: *p<0.1; **p<0.05; ***p<0.01. Robust standard errors in parentheses. Each column reports an OLS regression estimating the effect of being assigned to the *Public treatment* condition (i.e., public scrutiny) on two outcomes. In Columns (1) and (2), the outcome is modal rank belief (mrb), a measure of modesty, defined on a scale from 1 (highest self-assessed rank) to 18 (lowest). In Columns (3) and (4), the outcome is placement (true rank minus modal rank belief), a measure of confidence bias, defined on a scale from −17 to 17; higher values indicate greater overestimation of relative standing. Columns (2) and (4) control for gender, age, number of completed university years, number of languages spoken, an indicator for risk tolerance, a learning indicator based on changes in real-effort task scores across rounds, and three binary indicators for whether the participant reports majoring in STEM, Economics/Statistics, or Health Sciences.

Figure 3: Distributions of CATEs

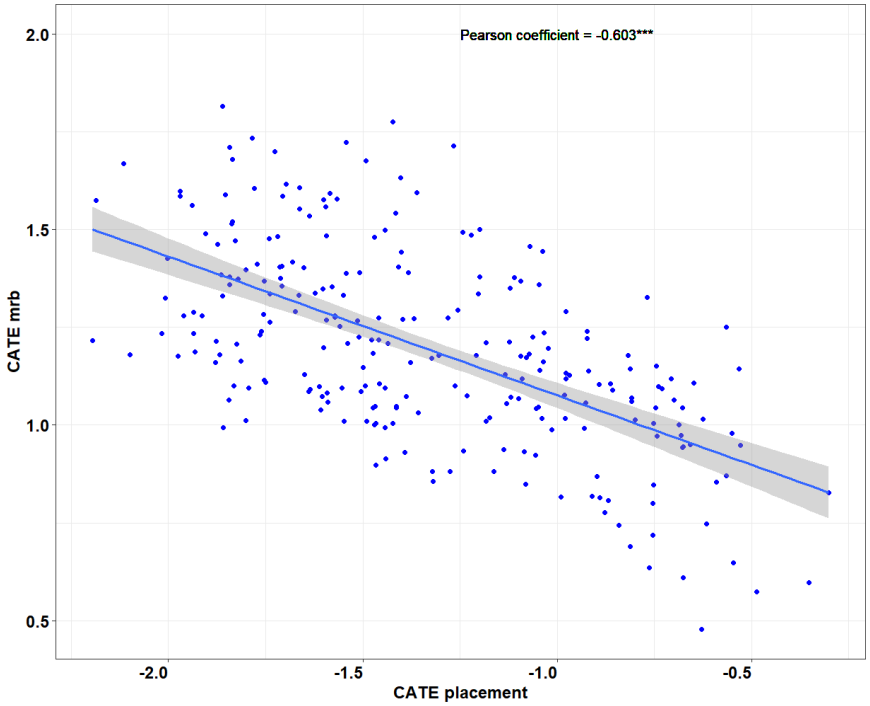


(a.) Modal rank belief (mrb)

(b.) Belief error (placement)

NOTES: The figure shows the distribution of CATEs estimated on the two measures of self-consciousness: mrb ("modesty") and placement ("confidence bias"), and their summary statistics. Dashed vertical lines indicate the mean.

Figure 4: Correlation of CATEs for two measures of self-consciousness



NOTES: The figure shows a scatterplot and linear regression fit between the CATEs estimated on the two measures of self-consciousness: mrb ("modesty") and placement ("confidence bias"). The text reports the Pearson correlation coefficient and its statistical significance (1%).

Section E in the Appendix reports correlations between individual CATEs and pre-treatment characteristics. Two key patterns emerge. First, susceptibility to self-consciousness under public scrutiny is not correlated with being enrolled in a health-related major, suggesting that health expertise does not systematically affect emotional responsiveness in our setting. This is consistent with the fact that our *Public treatment* is not focused on health, implying conservative estimates of the correlation between CATEs and vaccine attitudes. Second, few covariates show statistically significant associations with CATEs. The most consistent correlate is age, which also maps onto the number of completed university years. In contrast, there is little evidence that field of study or task learning (as measured by improvements in real-effort task scores over successive experimental rounds) strongly predicts individual treatment effects. One exception is the CATE for confidence bias, which is negatively associated with having studied game theory in one's university coursework. These participants exhibit less confidence bias following treatment, potentially reflecting greater familiarity with laboratory experiments or strategic reasoning. Overall, however, the significant correlations are modest in size—approximately half a standard deviation of the CATE distribution.

4.2 Module 2: Correlating Self-Consciousness intensity and Health Attitudes

The correlations between individual treatment effects (CATEs) and vaccine-related attitudes are reported in Table 3. Although statistical significance varies due to the limited sample size, the direction of the coefficients is consistent across both measures of self-consciousness—modesty and confidence bias—when interpreted in terms of the strength of individual responses to public scrutiny. This suggests that public scrutiny does not causally affect vaccine attitudes, but that individual susceptibility to self-consciousness instead correlates with pre-existing misbeliefs.

If the correlation were driven by a causal effect of public scrutiny, we would expect individuals with stronger self-consciousness responses to be less willing to express controversial views and to demand higher compensation to reveal them. We find the opposite: participants with larger absolute CATEs (i.e., stronger responsiveness to public scrutiny) are more likely to endorse vaccine misbeliefs and are less preoccupied with concealing them. By contrast, those with smaller absolute CATEs (i.e., individuals less emotionally affected by public scrutiny) express more accurate beliefs and are more interested in keeping their beliefs private.

Among the outcomes, statistical significance is achieved only for the belief that vaccines can cause the illness they are meant to prevent (Statement 2). Compared to the misbelief about adverse effects, this item may provide a clearer test of misinformation: it is objectively false and less value-laden, making it easier to measure cleanly with a Likert scale. The magnitude is meaningful: agreement with this misbelief is associated with a one standard deviation increase in the CATE for modesty (0.25), and about half that size for confidence bias. These results are confirmed—and strengthened in the case of the misbelief that vaccines cause illness—when using an alternative and stricter definition of agreement (values above 6 on the 11-point Likert scale); results are presented in Table G.4 of the Appendix. The direction of correlation is further confirmed by the misbelief sum score; however, this variable does not display statistical significance, as it reflects the cumulative contribution of all three misbelief systems—of which only illness causation displays significance—and the aggregation issues already discussed in Section 3.4.

We interpret these results as suggestive evidence that individuals who are more susceptible to self-consciousness also tend to hold more sceptical views about vaccines. In other words, the observed misbeliefs appear to reflect pre-treatment attitudes rather than being caused by the elicitation of

self-consciousness through public scrutiny. Notably, these individuals do not seem more reluctant to share their attitudes: instead, those who exhibit stronger responses to self-consciousness are just as likely, or even more likely, to express controversial views.

This interpretation is further supported by the relationship between confidence bias and health-secrecy.[23] Participants with higher CATEs for confidence bias—reflecting heterogeneous responses to public scrutiny along the placement dimension—are also more likely to request very high compensation to make their vaccine attitudes public. Because vaccine attitudes are self-reported, they may partly reflect concerns about disclosure; we therefore interpret this evidence in conjunction with our separate measure of health-secrecy. If self-consciousness elicitation had a causal effect on health-secrecy, we would expect individuals exhibiting stronger responses to public scrutiny along this dimension to be more reluctant to disclose their attitudes. Instead, we observe a positive correlation, consistent with the interpretation that differences in disclosure behaviour reflect underlying heterogeneity rather than treatment-induced concealment.

In addition, we find that participants with higher levels of vaccine scepticism tend to overestimate the extent to which others hold similar misbeliefs. As shown in Figure 5, there is a strong and statistically significant correlation between a participant's own scepticism score and their overestimation of the group's average score (Pearson coefficient $= 0.548$, $p < 0.01$). This pattern is consistent with a false consensus effect: individuals who endorse misbeliefs project those views onto others, normalising their beliefs through perceived social support. Such a pattern may contribute to the persistence of vaccine hesitancy by reducing the perceived deviation from prevailing social norms.

This interpretation aligns with existing findings that link vaccine hesitancy to conspiracy beliefs and anti-intellectualism (e.g., van Prooijen, 2018; van Prooijen and Böhm, 2024; Callaghan et al., 2019; Farhart et al., 2022; Gagliardi, 2025), and with recent evidence on conspiracy-related thinking in the context of COVID-19 (Miller, 2020; Imhoff and Lamberty, 2020). Taken together, our results are consistent with a broader predisposition toward being manipulated—both emotionally and informationally—which may help contextualise the observed alignment between self-consciousness responsiveness and vaccine scepticism, in line with evidence that conspirational mindset is associated with several cognitive biases (for a recent review, see Gagliardi, 2025).

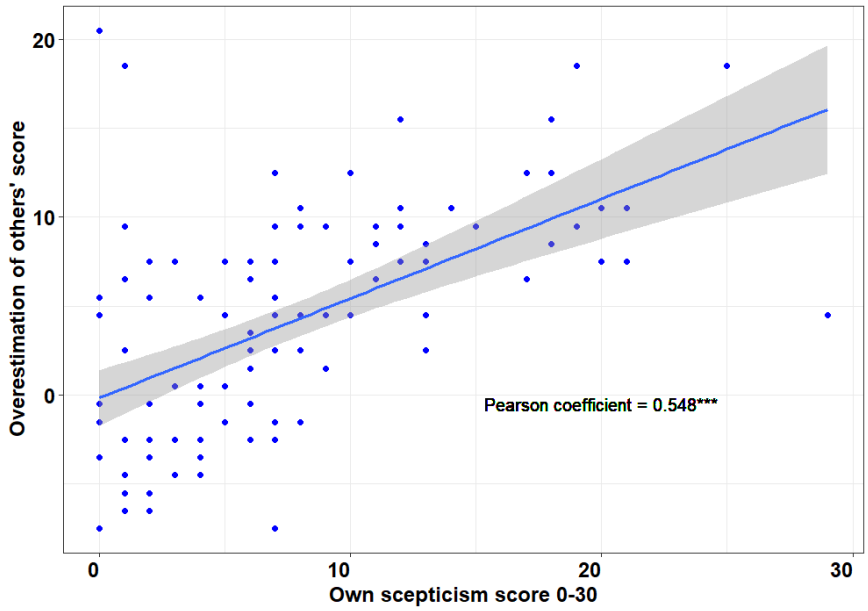
[23] For the placement (confidence-bias) measure, the CATEs are directional, with larger absolute values associated with more negative placement adjustments in response to public scrutiny.

Table 3: Correlation of CATEs and health attitudes

Health attitude measure	Coefficient	s.e.
Correlation with CATE for mrb ("modesty")		
Misbelief: illness causation	0.253*	(0.137)
Misbelief: adverse effects	0.080	(0.163)
Aggregate misbelief: sum score (0-30)	1.690	(2.248)
High health-secrecy	-0.278	(0.180)
Correlation with CATE for placement ("confidence bias")		
Misbelief: illness causation	-0.148*	(0.076)
Misbelief: adverse effects	-0.074	(0.090)
Aggregate misbelief: sum score (0-30)	-0.885	(1.242)
High health-secrecy	0.201**	(0.098)
N	112	

NOTES: Correlations are estimated by bivariate OLS models between the CATEs (individual causal effects of self-consciousness elicitation that embed covariates in their estimation) and different health attitude measures. "High health-secrecy" equals 1 for individuals who demand more than 45 euros to make their health beliefs public. The reduced sample size relative to Table 1 reflects that only treated individuals were surveyed on health attitudes. *p<0.1; **p<0.05; ***p<0.01.

Figure 5: Own vs. perceived vaccine scepticism



NOTES: Each dot represents one participant. The x-axis plots participants' own vaccine scepticism score (0-30), and the y-axis plots the difference between their estimate of others' average scepticism and the true sample mean. A positive value on the y-axis indicates overestimation. The blue line shows a linear fit with 95% confidence interval. Pearson correlation = 0.548 (*** $p < 0.01$).

4.3 Limitations and Policy Implications

While this paper has a number of limitations, we believe the results offer insights that could drive future research in the understudied and policy-relevant topic of doctor-patient interactions, for the promotion of preventive healthcare. Some limitations stem from the fact that Module 2, which presented subjects with questions on health attitudes, was nested within a larger, pre-existing experimental study. In particular, this had three implications. Because Module 2 was added while the larger study was already underway, the corresponding analysis is exploratory, meaning that it was not pre-registered, and we were not able to secure additional funding to incentivise the revelation of health-secrecy. However, vaccine attitudes are generally not thought to be influenced by monetary incentives.

Moreover, because Module 2 was only added in experimental sessions comprising treated subjects, our empirical analysis does not include control units and therefore does not neatly identify the causal effect of self-consciousness on health attitudes. Instead, it shows how susceptibility to self-consciousness correlates with health attitudes. We believe that our observed results mitigate this design concern: indeed, they suggest the *absence of a causal relationship*, and indicate, as a relevant avenue for future research, a possible positive correlation between a conspiratorial mindset, potentially characterised by higher vaccine hesitancy, and greater susceptibility to poor doctor-patient communication. In other words, those who may become more self-conscious when exposed to disparaging comments from doctors are also those who hold negative views on vaccines, making policy interventions more salient.

In this regard, another limitation is that we conducted a laboratory experiment rather than addressing our question in the field. Indeed, our results are based on a limited sample of Italian university students, who are young and generally more educated than the general population. Older and less educated individuals tend to be more responsive to vaccine communication campaigns (e.g., see Dominici et al. (2025) for the flu vaccine in Italy and Dominici and Dahlström (2025) for an HPV vaccine campaign in Sweden). In our data, the magnitude of individual treatment effects (CATEs) is positively correlated with age. While we cannot guarantee that the results apply to the general population or to countries with different health policies, these elements suggest that, if anything, our estimates may be conservative. Future studies in the field could also aim to disentangle actual vaccine attitudes from preferences over revealing them, which in our setting are

elicited through separate questions. While this approach already provides partial insight, it still relies on self-reported data; a clearer disentanglement would require comparing revealed behaviour in the field with survey responses. More generally, our main contribution lies in highlighting the co-occurrence of susceptibility to self-conscious emotions and vaccine hesitancy as an avenue for future research that should be pre-registered, replicable, and ideally conducted in field settings or in health-related contexts that more directly elicit self-conscious emotions.

Another consideration concerns the way our main measures of self-consciousness are constructed from the belief-elicitation task. In the literature, belief-elicitation methods vary considerably and can differ in the precision, interpretability, and behavioural meaning of the measures they produce (e.g. Schlag and van der Weele, 2015; Charness et al., 2021). We selected our measures for three reasons: they richly capture beliefs about relative position, they are directly comparable to the actual rank, and they are comprehensible to the principal in the treatment condition. Nevertheless, we recognise that the behavioural response we capture may reflect a combination of self-conscious adaptation to public exposure and "other-regarding" considerations toward the principal's payoff, or, more generally, strategic adjustments that work in the same countervailing direction. As such, our estimates should be interpreted as conservative, or lower-bound, measures of the pure self-consciousness response, implying that in applied settings—such as doctor-patient communication—any missteps that inadvertently induce self-conscious emotions (such as shame) could have a stronger behavioural impact than our estimates suggest.

A further possibility is that sensitivity to public scrutiny may not only represent a situational self-conscious response but may also reflect other underlying traits. For example, low self-esteem or identification with alternative social norms might predispose individuals to adopt belief systems that provide compensatory anticipatory utility, such as alternative truths or conspiracy theories. Although our design cannot disentangle these deeper mechanisms, this interpretation aligns with our findings in Module 2 and highlights an important avenue for future research.

While our correlations show a consistent pattern across different measures of both self-consciousness and health attitudes, they are based on a small sample, resulting in large variances and low statistical power. At the same time, the brevity of Module 2 relative to Module 1, in our setting, restricts our ability to pinpoint specific emotional mechanisms behind self-consciousness. Our evidence is not conclusive, and a complete mechanism analysis is beyond the scope of our exploratory anal-

ysis, but it sheds light on an important psychological dynamic with potentially significant policy implications, suggesting this field as a relevant avenue for more structured future research.

5 Conclusion

This paper contributes to a growing literature on the emotional underpinnings of vaccine scepticism by exploring how individuals' susceptibility to self-consciousness correlates with their health attitudes. Using a randomised laboratory experiment, we estimate both average and individual-level effects of public scrutiny on the emergence of self-conscious emotions, measured in terms of rank beliefs. We then link these responses to a follow-up survey on vaccine-related misbeliefs and the willingness to conceal them. We find that individuals who are more emotionally responsive to self-consciousness—measured through changes in self-assessed relative rank and confidence bias—are also more likely to endorse misbeliefs about vaccines. Importantly, this self-consciousness was not evoked in a health context, but through a task unrelated to health attitudes. This makes the observed correlation particularly striking, and likely conservative: if we had induced self-consciousness more directly tied to health beliefs, the alignment may have been even stronger.

Furthermore, the experimental context reflects a one-time interaction with an anonymous observer and includes only university students. Participants had no ongoing relationship with the principal evaluating them, and their high education level and young age position them among the least responsive to health communication nudges, according to the existing literature. In real-world settings—characterised by a wider population range and interactions with family doctors or community health workers—susceptibility to self-consciousness may be heightened by the perceived relational or reputational cost. Our findings, therefore, raise questions about how self-conscious individuals communicate with others about health topics and whether these dynamics contribute to disengagement from medical conversations. We interpret our results as correlational rather than causal; however, the pattern is robust across two distinct measures of self-consciousness and multiple belief outcomes. The results suggest that susceptibility to self-consciousness may be a behavioural marker of deeper psychological traits—such as defensiveness or suggestibility—that also shape resistance to scientific information.

Taken together, these findings underscore a policy challenge made especially salient by the dan-

ger of future virus-borne pandemics, for which vaccines represent the main protection for public health. Individuals who are most susceptible to self-consciousness may simultaneously hold the strongest misbeliefs and be the least willing to seek guidance from medical professionals. Communication strategies that inadvertently evoke self-conscious emotions—such as authoritative or dismissive tones—may therefore risk having counterproductive effects on vaccine-hesitant populations. Future research should examine manifestations of emotional responses to health messaging across different population segments and consider how self-consciousness interacts with identity, trust, and the perceived credibility of the messenger, thereby providing policymakers with further evidence, based on pre-registration, that supports replicability and a thorough analysis of the psychological mechanisms that inform behavioural interventions in the health domain.

Declaration of competing interest

The authors declare that they have no financial or personal relationships that could be perceived as having influenced the research presented in this paper.

References

C. Alamaa. Preregistration: The role of the observer for self-assessment—gender differences in image-concerns. Open Science Framework, 2023, June 11. URL https://doi.org/10.17605/OSF.IO/8Z39J. Preregistration. OSF registration ID: 8Z39J.

C. Alamaa. *Belief Exposure and Economic Behaviour: Experimental Evidence on the Role of Feedback, Social Evaluation, and Gender in Health Attitudes and Labour Market Outcomes*. Department of Economics, Stockholm University, 2025. URL https://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Asu%3Adiva-248965.

M. Alsan and S. Eichmeyer. Experimental Evidence on the Effectiveness of Non-Experts for Improving Vaccine Demand. *American Economic Journal: Economic Policy*, 16.1:394–414, 2024.

J. Andreoni and B. D. Bernheim. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636, 2009.

A. Armand, B. Augsburg, A. Bancalari, and K. K. Kameshwara. Religious proximity and misinformation: Experimental evidence from a mobile phone-based campaign in india. *Journal of Health Economics*, 96:102883, 2024. doi: 10.1016/j.jhealeco.2024.102883.

N. Ashraf, E. Field, and J. Lee. Household bargaining and excess fertility: an experimental study in zambia. *American Economic Review*, 104(7):2210–2237, 2014.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

S. Athey and S. Wager. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5 (2):37–51, 2019.

K. Barron, M. T. Damgaard, and C. Gravert. Nudge me! a field experiment on reminders for medication adherence. *Journal of Economic Behavior & Organization*, 241:107368, 2026. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2025.107368. URL https://www.sciencedirect.com/science/article/pii/S0167268125004858.

C. Betsch, P. Schmid, D. Heinemeier, L. Korn, C. Holtmann, and R. Böhm. Beyond confidence: Development of a measure assessing the 5c psychological antecedents of vaccination. *PLOS ONE*, 13(12):e0208601, 2018.

P. Bordalo, N. Gennaioli, and A. Shleifer. Memory, attention, and choice. *The Quarterly journal of economics*, 135 (3):1399–1442, 2020.

R. Bénabou and J. Tirole. Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678, 2006.

T. Callaghan, M. Motta, S. Sylvester, K. L. Trujillo, and C. C. Blackburn. Parent psychology and the decision to delay childhood vaccination. *Social Science & Medicine*, 238:112407, 2019.

G. Charness, U. Gneezy, and V. Rasocha. Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256, 2021. doi: 10.1016/j.jebo.2021.06.019.

D. L. Chen, M. Schonger, and C. Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2015.12.001. URL https://www.sciencedirect.com/science/article/pii/S2214635016000101.

D. Danz, L. Vesterlund, and A. J. Wilson. Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–83, September 2022. doi: 10.1257/aer.20201248. URL https://www.aeaweb.org/articles?id=10.1257/aer.20201248.

L. Dolezal and B. Lyons. Health-related shame: an affective determinant of health? *Medical Humanities*, 43(4):257–263, 2017. ISSN 1468-215X. doi: 10.1136/medhum-2017-011186. URL https://mh.bmj.com/content/43/4/257.

A. Dominici and L. A. Dahlström. Targeting vaccine information framing to recipients' education: A randomized trial. *Health Economics*, 34(12):2317–2337, 2025. doi: https://doi.org/10.1002/hec.70036. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.70036.

A. Dominici, F. Panizza, E. Bilancini, and L. Boncinelli. The limits and perils of gentle communication against vaccine hesitancy: an informational trial. *Working Paper*, 2025. URL https://drive.google.com/file/d/1hv7ZKLlY8Gxn2qWicGr-0Vcng-iePpE9/view.

P. Dupas. Health behavior in developing countries. *Annual Review of Economics*, 3:425–449, 2011.

T. Ellingsen and M. Johannesson. Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008, 2008.

N. M. Else-Quest, A. Higgins, C. Allison, and L. C. Morton. Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological Bulletin*, 138(5):947–981, 2012. doi: 10.1037/a0027930.

A. Falk and A. Ichino. Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57, 2006.

C. E. Farhart, E. Douglas-Durham, K. L. Trujillo, and J. A. Vitriol. Vax attacks: How conspiracy theory belief undermines vaccine support. *Progress in Molecular Biology and Translational Science*, 188(1):135–169, 2022. doi: 10.1016/bs.pmbts.2021.11.004.

U. Fischbacher and F. Föllmi-Heusi. Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547, 2013.

L. Gagliardi. The role of cognitive biases in conspiracy beliefs: A literature review. *Journal of Economic Surveys*, 39 (1):32–65, 2025. doi: 10.1111/joes.12604.

B. Greiner. Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125, 2015. URL https://EconPapers.repec.org/RePEc:spr:jesaex:v:1:y:2015:i:1:d:10.1007_s40881-015-0004-4.

Z. Grossman and J. van der Weele. Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217, 2017.

C. R. Harris and R. S. Darby. Shame in physician–patient interactions: Patient perspectives. *Basic and Applied Social Psychology*, 31(4):325–334, 2009.

R. Imhoff and P. Lamberty. A bioweapon or a hoax? the link between distinct conspiracy beliefs about the coronavirus disease (covid-19) outbreak and pandemic behavior. *Social Psychological and Personality Science*, 11(8):1110–1118, 2020.

M. A. Jaeb and K. E. Pecanac. Shame in patient-health professional encounters: A scoping review. *International Journal of Mental Health Nursing*, 33(5):1158–1169, 2024. doi: 10.1111/inm.13323. URL https://onlinelibrary.wiley.com/doi/full/10.1111/inm.13323.

A. Kata. A postmodern pandora's box: anti-vaccination misinformation on the internet. *Vaccine*, 28(7):1709–1716, 2010.

J. M. Miller. Do COVID-19 conspiracy theory beliefs form a monological belief system? *Canadian Journal of Political Science/Revue canadienne de science politique*, 53(2):319–326, 2020. doi: 10.1017/S0008423920000337.

K. M. Norgaard. "we don't really want to know": Environmental justice and socially organized denial of global warming in norway. *Organization & Environment*, 19(3):347–370, 2006.

K. Nyborg. I don't want to hear about it: Rational ignorance among duty-oriented consumers. *Journal of Economic Behavior & Organization*, 79(3):263–274, 2011.

G. D. Salali and M. S. Uysal. Covid-19 vaccine hesitancy is associated with beliefs on the origin of the novel coronavirus in the uk and turkey. *Psychological Medicine*, pages 1–3, 2020.

P. Sankar and N. L. Jones. To tell or not to tell: primary care patients' disclosure deliberations. *Archives of Internal Medicine*, 165(20):2378–2383, 2005. doi: 10.1001/archinte.165.20.2378. URL https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/486598.

K. H. Schlag and J. J. van der Weele. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18:457–490, 2015. doi: 10.1007/s10683-014-9416-x.

A. Schotter and I. Trevino. Belief elicitation in the laboratory. *Annual Review of Economics*, 6:103–128, 2014.

J. P. Tangney, J. Stuewig, and D. J. Mashek. Moral emotions and moral behavior. *Annual Review of Psychology*, 58: 345–372, 2007. doi: 10.1146/annurev.psych.56.091103.070145.

J. P. E. Tangney and K. W. Fischer. Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride. In *The idea for this volume grew out of 2 pivotal conferences. The 1st conference, on emotion and cognition in development, was held in Winter Park, CO, Sum 1985. The 2nd conference, on shame and other self-conscious emotions, was held in Asilomar, CA, Dec 1988.* Guilford Press, 1995.

A. Tesser. Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology*, 21:181–227, 1988.

R. L. Thornton. The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–1863, 2008.

J. L. Tracy and R. W. Robins. The nature of pride. *The self-conscious emotions: Theory and research*, pages 263–282, 2007.

S. T. Trautmann and G. van de Kuilen. Belief elicitation: A horse race among truth serums. *Economic Journal*, 125 (589):2116–2135, 2015.

J.-W. van Prooijen. Architecture of belief. In *The Psychology of Conspiracy Theories*, chapter 3, pages 38–39. Routledge, London, first edition, 2018. doi: 10.4324/9781315525419-3.

J.-W. van Prooijen and N. Böhm. Do conspiracy theories shape or rationalize vaccination hesitancy over time? *Social Psychological and Personality Science*, 15(4):421–429, 2024. doi: 10.1177/19485506231181659.

H. Wang, J.-W. van Prooijen, and P. A. van Lange. How perceived coercion polarizes unvaccinated people: The mediating role of conspiracy beliefs. *Journal of Health Psychology*, 30(9):2354–2367, 2025.

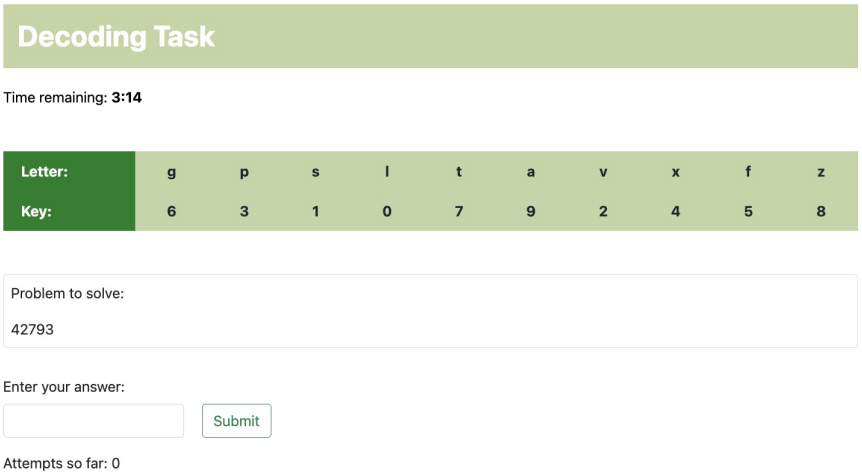
WHO. Framework on integrated, people-centred health services. Technical report, April 2016. URL https://apps.who.int/gb/ebwha/pdf_files/WHA69/A69_39-en.pdf.

APPENDIX

A Experimental Module 1: Real-Effort Task and Rank-Elicitation Task

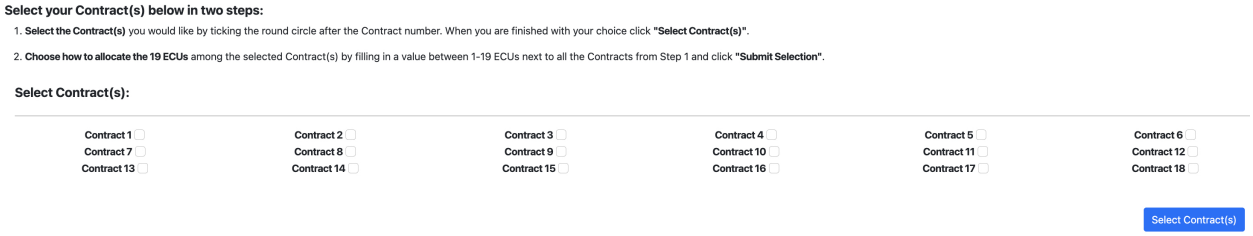
Real-Effort Task. Each correctly solved problem is rewarded with one point, and there are no punishment for wrong answers. The clock in the top-left corner is a timer counting down from the allotted time duration of 4:00 minutes. The Decoding-Task involves *decoding* the 5-digit number, using the Key to a 5-letter string that should be inserted in the answer-box. In the example the problem is to decode "42793" as 4 to X; 2 to V; 7 to T; 9 to A; and 3 to P and submit "XVTAP" as an answer. The answer was not case-sensitive.

Figure A.1: Real-Effort Task: Example Screen



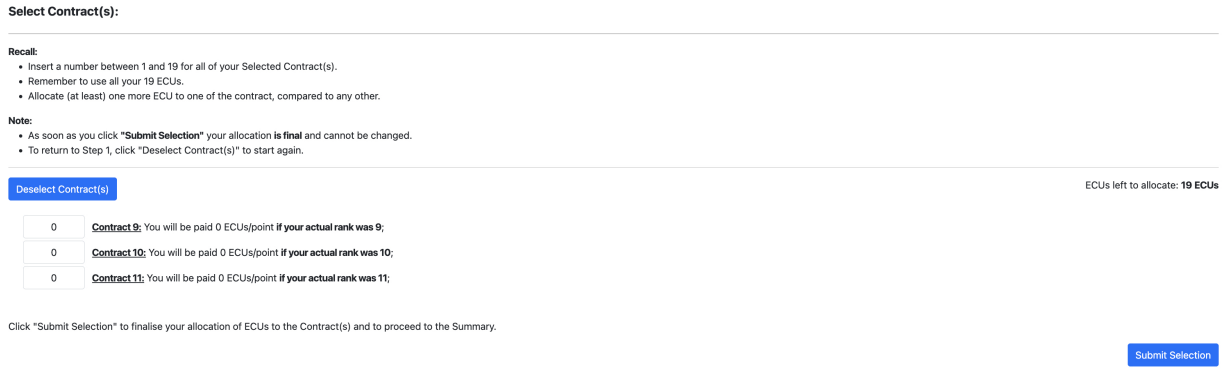
NOTES: The figure shows an example screen of the real-effort task used in all three rounds in Module 1.

Figure A.2: Rank-Selection Task - Step A: Example Screen



NOTES: The figure shows an example-screen of the first step of the rank-investment selection, in which agents select which ranks they would like to allocate any endowment larger thank zero.

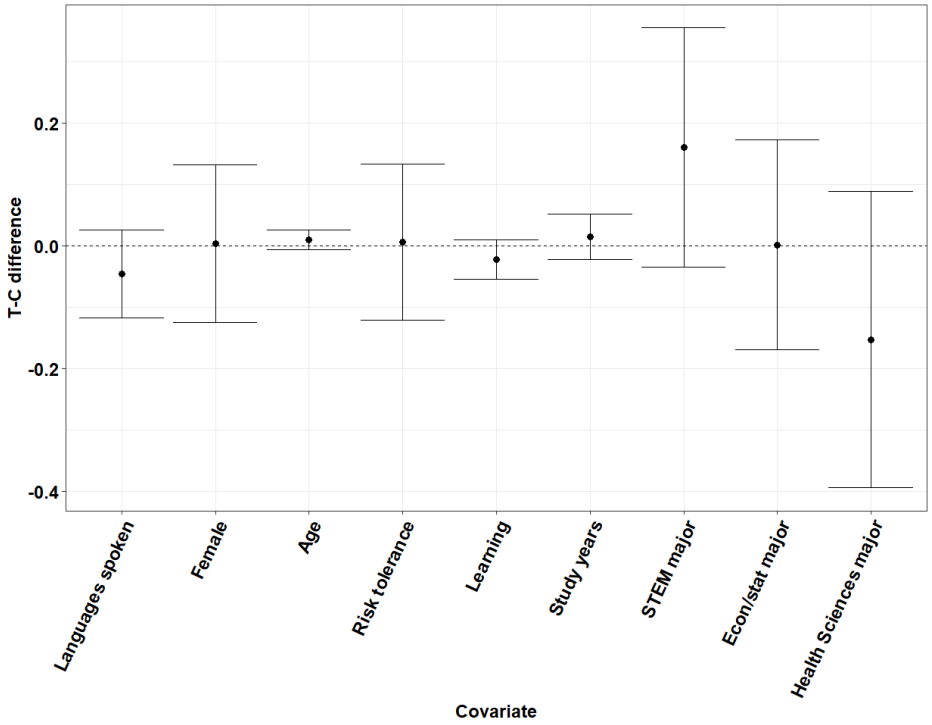
Figure A.3: Rank-Selection Task - Step B: Example Screen



NOTES: The figure shows an example-screen of the second step of the rank-investment selection, in which agents select how much they would like to allocate to each of the selected ranks from step 1 (Step A).

B Balance

Figure B.4: Covariates' balance



NOTES: The figure shows OLS coefficients obtained by regressing, in the full sample, the binary *public treatment* indicator on all covariates (i.e., conditional balance).

C Summary statistics

Table C.1: Covariates in the full sample: summary statistics

	Scale	Mean	St.dev	Min	Max
Female	0/1	0.514	0.501	0	1
Age	18–50+	24.873	4.462	18	51
Higher educ. (yrs. completed)	0–6	3.348	1.823	0	6
No. of spoken languages	1–7	3.143	0.91	2	7
Risk-tolerance	0/1	0.551	0.498	0	1
Learning (*real-effort task*)	-4–9	1.037	1.999	-4	9
Study-field majors					
STEM	0/1	0.131	0.338	0	1
Econ./Stat.	0/1	0.22	0.415	0	1
Health/Medicine	0/1	0.078	0.268	0	1

NOTES: The full sample described in this table is used to compute the overall ATE of self-consciousness elicitation and individual CATEs.

Figure C.5: Misbelief illness: full distribution

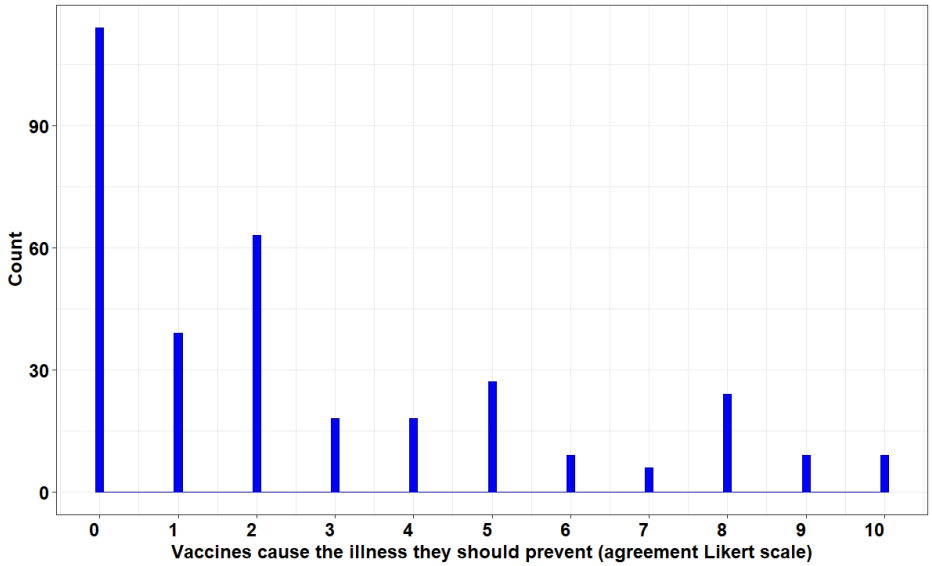


Figure C.6: Misbelief weakened immune system: full distribution

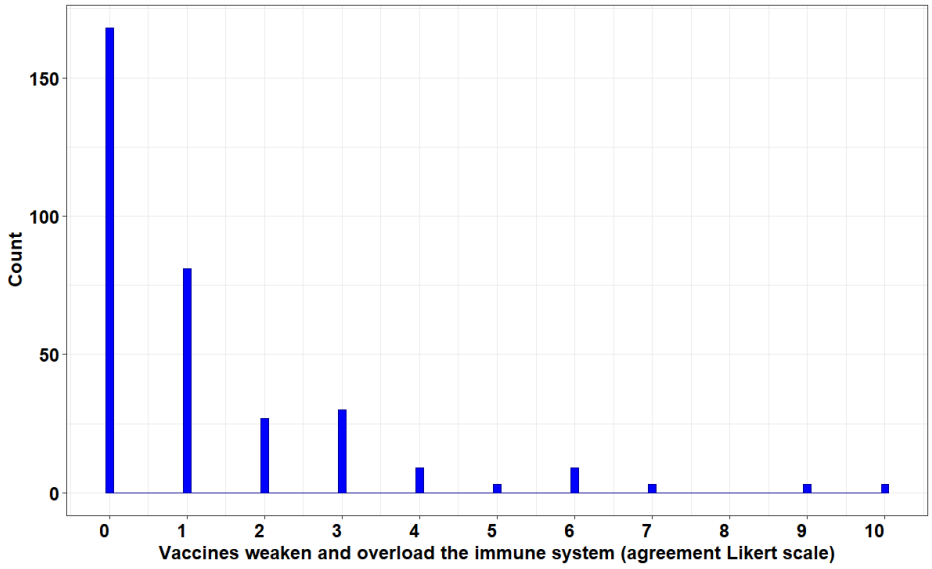


Figure C.7: Misbelief adverse effects: full distribution

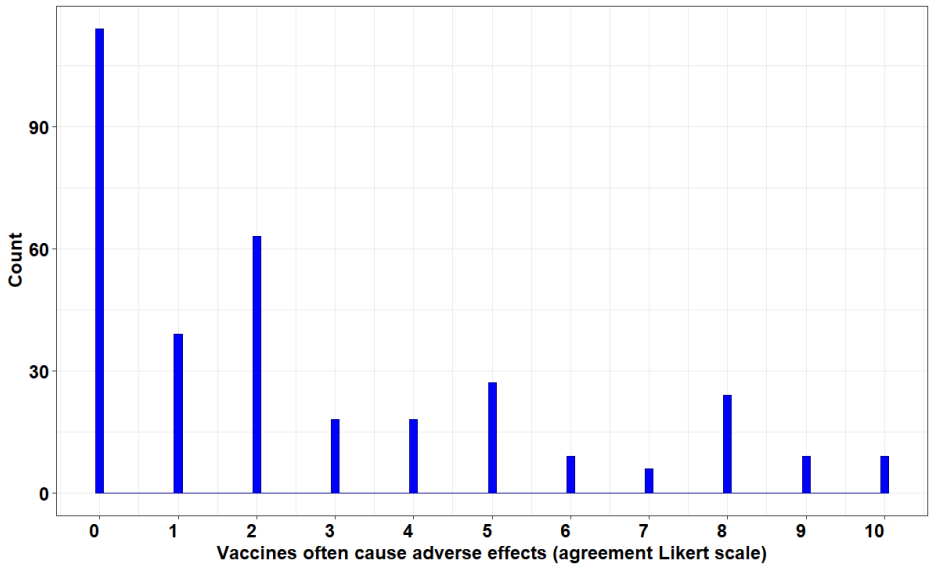


Table C.2: Covariates and outcomes among the treated: summary statistics

	Mean	St.dev	Min	Max
Covariates				
N. of languages	3.143	0.91	2	7
Years of university	3.348	1.823	0	6
Female (dummy)	0.514	0.501	0	1
Age	24.873	4.462	18	51
Risk aversion score (dummy)	0.551	0.498	0	1
STEM studies (dummy)	0.131	0.338	0	1
Econ/stat studies (dummy)	0.22	0.415	0	1
Medical studies (dummy)	0.078	0.268	0	1
Learning	1.037	1.999	-4	9
Health attitudes outcomes				
Agrees that vaccines weaken the immune system [0-10]	1.214	1.896	0	10
Agrees that vaccines cause illness [0-10]	2.634	2.935	0	10
Agrees that vaccines cause serious and permanent adverse effects [0-10]	3.661	3.074	0	10
Skepticism score (sum of agreements)	7.509	6.089	0	29
Guess of avg scepticism score	11.545	6.221	0	28
Health secrecy [ECUs]	291.795	198.333	0	500
Misbelief immune (agreem. immune>5)	0.054	0.226	0	1
Misbelief illness (agreem. illness>5)	0.17	0.377	0	1
Misbelief adverse effects (agreem. adverse>5)	0.259	0.44	0	1
Extreme Health shame (>75th percentile)	0.393	0.491	0	1

NOTES: Treated units described in this table are used to correlate the CATEs (individual effect of self-consciousness elicitation) and vaccine beliefs. Health secrecy is measured by asking how many Experimental Credit Units (ECU) each subject would require in order to make their aggregate vaccine misbelief score public to the principal. 10 ECU = 1 EUR.

D Causal forests

Causal forests are a supervised machine learning technique designed by Athey and Imbens (2016) to expand random forest in the context of causal inference. Originally intended for heterogeneity analysis, causal forests estimate individual causal effects exploiting observable individual-level covariates, opening a wide range of possibilities. In our case, we exploit the availability of individual causal estimates from the first stage of our analysis to overcome the absence of a control group in the second exploratory stage.

More formally, causal forests estimate Conditional Average Treatment Effects (CATE), allowing us to understand how the effects of a given intervention vary across different individuals. For a binary treatment scenario, the CATE is defined as:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid \boldsymbol{X_i} = \boldsymbol{x}]$$

Here, Y_{1i} and Y_{0i} denote the potential outcomes under treatment and control conditions, respectively, while $\boldsymbol{X_i}$ represents the vector of observable characteristics for individual i. Following the methodology outlined by Athey and Wager (2019), we implement an "honest" approach to causal forest estimation. Specifically, we split the sample into two equally sized parts: the first half is utilised to construct a forest of 10,000 trees, while the second half is reserved for CATE estimation.

Each tree within the forest partitions the covariate space recursively, aiming to maximise heterogeneity in treatment effects. The algorithm selects splitting variables based on their contribution to heterogeneity, employing a loss function that prioritises predictive accuracy. The final leaves—representing the smallest partitions—contain a minimum of five observations to ensure robustness. After constructing the forest, the second half of the sample is used to estimate CATEs based on the covariate importance identified during the tree-building process.

In practice, we use a sample of 244 subjects and estimate CATEs based on 9 covariates. To verify the absence of small leaves, we compute the effective number of observations contributing to each individual's CATE estimate (the effective neighbourhood size), along with the proportion of predictive weight coming from treated and control units, respectively. This is done using the estimated forest weights w_{ij}, which quantify how much each training unit j contributes to the

prediction for unit i. Table D.3 below reports the distributions of (i) the effective neighbourhood size, and (ii) the treated and control weight shares, for both CATEs (mrb "modesty" and placement "overconfidence"). It shows that effective sizes are consistently well above 100, and that the weight shares are centred around 0.5, indicating strong overlap between treated and control observations in each prediction neighbourhood.

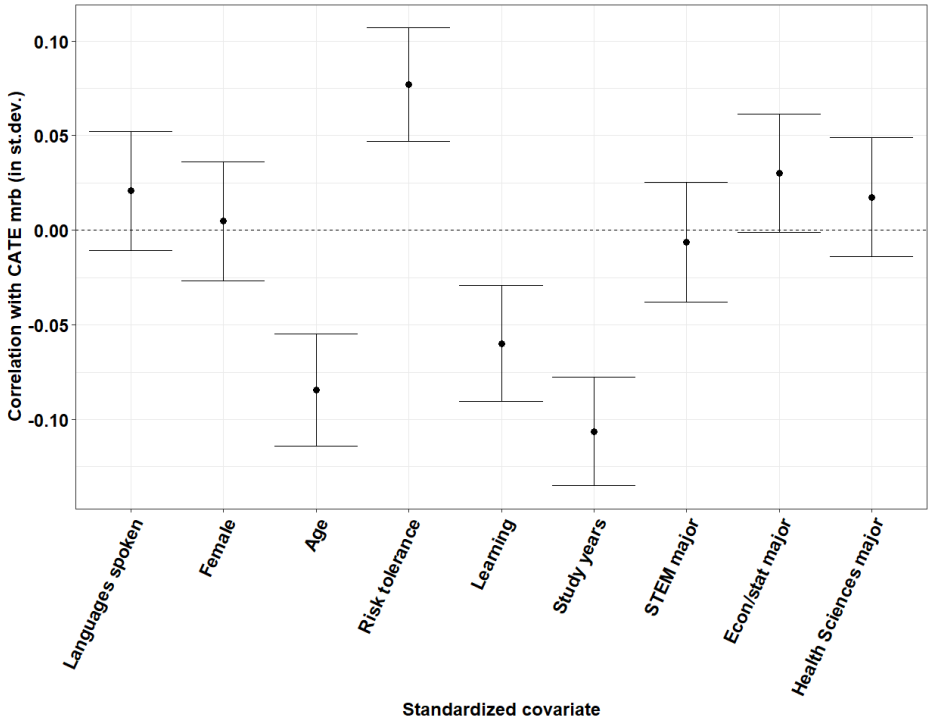
Table D.3: Effective neighbourhood sizes and treated/control weight shares

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Panel A: CATE for *mrb* ("modesty")						
Effective neighbourhood size (eff_n)	138.5	154.6	165.7	164.7	174.8	191.4
Treated weight share	0.4038	0.4428	0.4545	0.4579	0.4747	0.5259
Control weight share	0.4741	0.5253	0.5455	0.5421	0.5572	0.5962
Panel B: CATE for *placement* ("overconfidence")						
Effective neighbourhood size (eff_n)	137.3	153.9	166.2	164.4	174.8	195.4
Treated weight share	0.4080	0.4404	0.4523	0.4563	0.4709	0.5186
Control weight share	0.4814	0.5291	0.5477	0.5437	0.5596	0.5920

NOTES: The table summarises the distributions of sample sizes and treatment/control unit shares effectively used to compute CATEs. eff_n $= 1/\sum_j w_{ij}^2$, where w_{ij} is the out-of-bag prediction weight assigned to training unit j when estimating the Conditional Average Treatment Effect (CATE) for unit i.

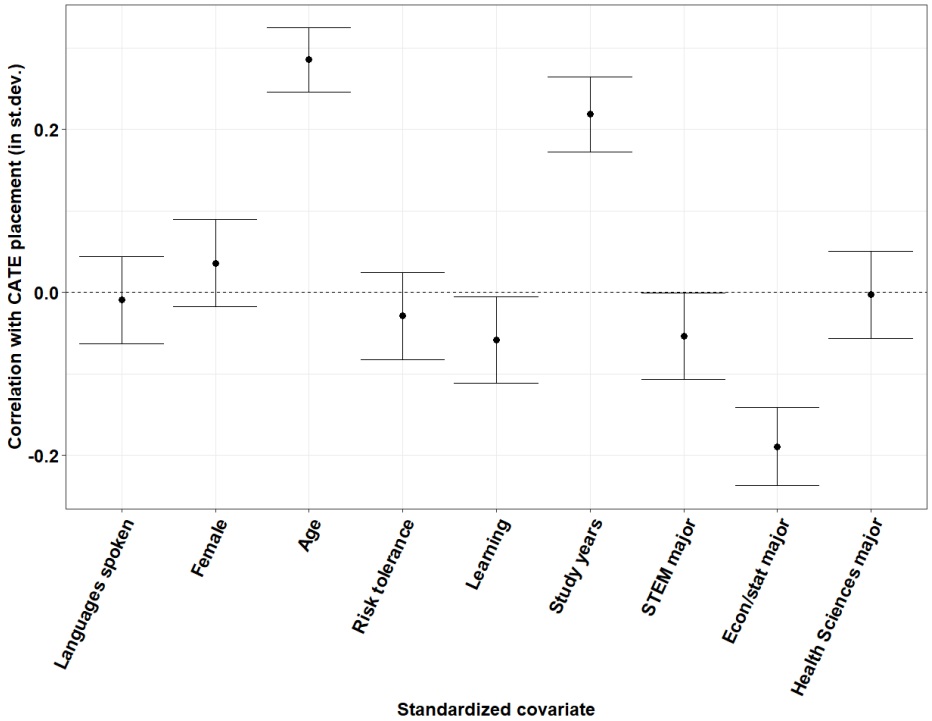
E CATEs by baseline characteristics

Figure E.8: CATE on mrb ("modesty") by baseline covariates



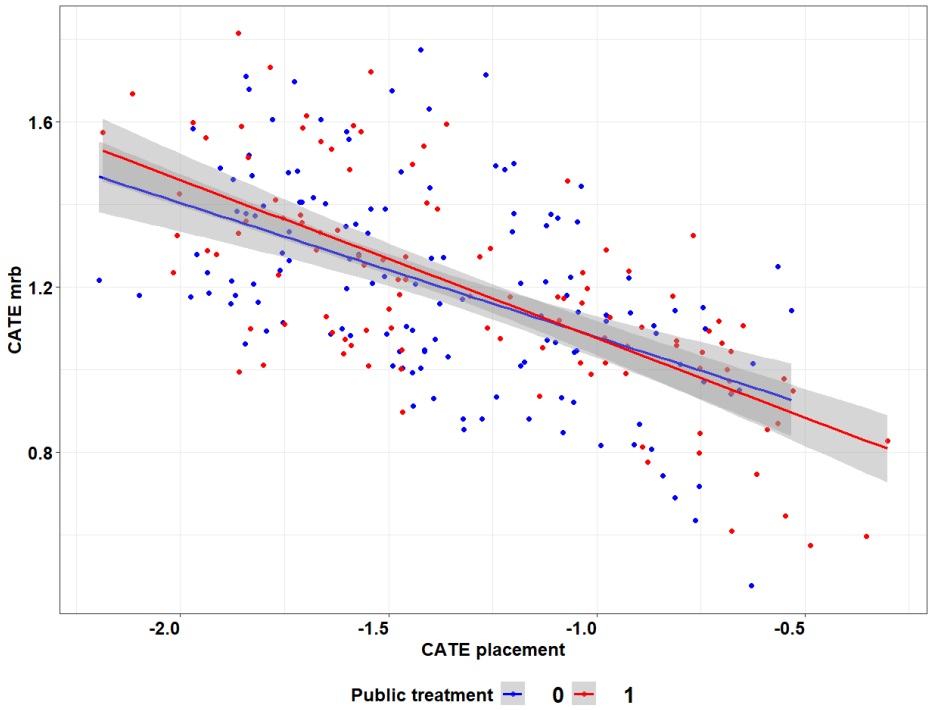
NOTES: The figure shows the correlation between the individual causal effect on mrb ("modesty") and baseline covariates, estimated by separate OLS regressions. Bars indicate 95% confidence intervals.

Figure E.9: CATE on placement ("confidence bias") by baseline covariates



NOTES: The figure shows the correlation between the individual causal effect on placement ("confidence bias") and baseline covariates, estimated by separate OLS regressions. Bars indicate 95% confidence intervals.

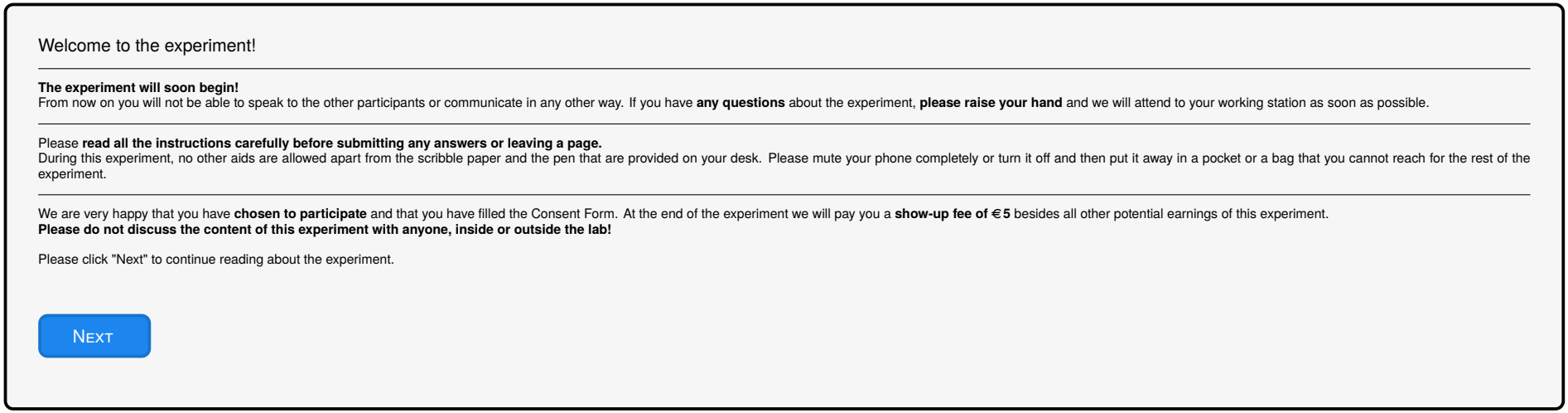
Figure E.10: CATEs by health outcomes observability



NOTES: The figures shows the value of the two CATEs (for mrb "modesty" and placement "confidence bias") colored by whether the individual is asked about vaccine beliefs (i.e., whether they were subject to the *Public* treatment and are present in Module 2 of the experiment).

F Experimental Instructions

Figure F.11: page 1



Welcome to the experiment!

The experiment will soon begin!
From now on you will not be able to speak to the other participants or communicate in any other way. If you have **any questions** about the experiment, **please raise your hand** and we will attend to your working station as soon as possible.

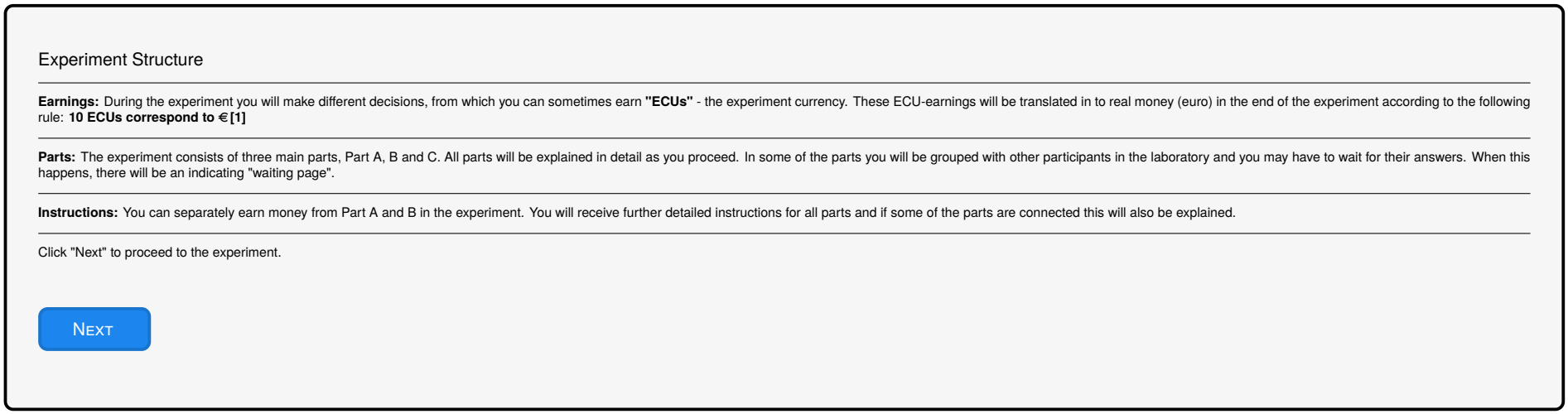
Please **read all the instructions carefully before submitting any answers or leaving a page.**
During this experiment, no other aids are allowed apart from the scribble paper and the pen that are provided on your desk. Please mute your phone completely or turn it off and then put it away in a pocket or a bag that you cannot reach for the rest of the experiment.

We are very happy that you have **chosen to participate** and that you have filled the Consent Form. At the end of the experiment we will pay you a **show-up fee of € 5** besides all other potential earnings of this experiment.
Please do not discuss the content of this experiment with anyone, inside or outside the lab!

Please click "Next" to continue reading about the experiment.

NEXT

Figure F.12: page 2



Experiment Structure

Earnings: During the experiment you will make different decisions, from which you can sometimes earn **"ECUs"** - the experiment currency. These ECU-earnings will be translated in to real money (euro) in the end of the experiment according to the following rule: **10 ECUs correspond to € [1]**

Parts: The experiment consists of three main parts, Part A, B and C. All parts will be explained in detail as you proceed. In some of the parts you will be grouped with other participants in the laboratory and you may have to wait for their answers. When this happens, there will be an indicating "waiting page".

Instructions: You can separately earn money from Part A and B in the experiment. You will receive further detailed instructions for all parts and if some of the parts are connected this will also be explained.

Click "Next" to proceed to the experiment.

NEXT

Figure F.13: page 3

Introductory questions

Select your age in years:

Use drop-down list with year span: *18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32; 33; 34; 35; 36; 37; 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 50+*

Gender:

◯ Man ◯ Woman

Select your main field of study (select "none" if you never been a student; select "other" and specify if your main study area is not in the list):

Drop-down list with educational tracks - List in two levels, where first level is not electable.
Humanities and Social Sciences: Anthropology/Archaeology; History; Linguistics and languages; Philosophy; Religion; The arts; Economics; Geography; Interdisciplinary studies; Political science; Psychology; Sociology
Natural Sciences: Biology; Chemistry; Earth science; Physics; Space science/Astronomy
Formal Sciences: Computer science; Logic; Mathematics/Statistics; System science
Professions and Applied Sciences: Agriculture; Architecture and design; Business; Education; Engineering and technology; Environmental studies and forestry; Journalism/media studies/ communication; Law; Library and museum studies; Medicine; Military sciences; Public administration/Public policy; Social work; Transportation
Other: Other; None

Other:
Specify:

Pop-up field if selected "other". If "none" selected, impossible to select any years of education.

Number of finished years:

◯ 0 ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 5+

Please tick the languages that you speak apart from Italian and indicate your level of command:

◯ English	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Spanish	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ French	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ German	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Albanian	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Other	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic
◯ Other	◯ Native ◯ Fluent ◯ Very Good ◯ Good ◯ Basic

Use Fill-in for Other. Possible to tick more than one. Use 5 tick-boxes + 2 others with free text.

Click "Next" to start **Part A** of the experiment.

NEXT

Figure F.14: page 4

Overview of Part A

Part A consists of two main tasks:

- A **"Decoding Task"**: solve problems for **4 minutes.**
- A **"Contract Selection":** choose a working contract that determines the pay you get for your correctly solved problems.

This is repeated in 3 rounds.

Part A has two roles: All participants will randomly be assigned one of the two **roles:** "Employee" or "Principal".

- **Employees:** perform 3 rounds of Decoding Tasks and Contract Selections.
- **Principals:** will be randomly **matched with 3 Employees**. The Principals will only perform the Decoding Task in Round 1.

Your assigned role shows up on your screen before the first Contract Selection.

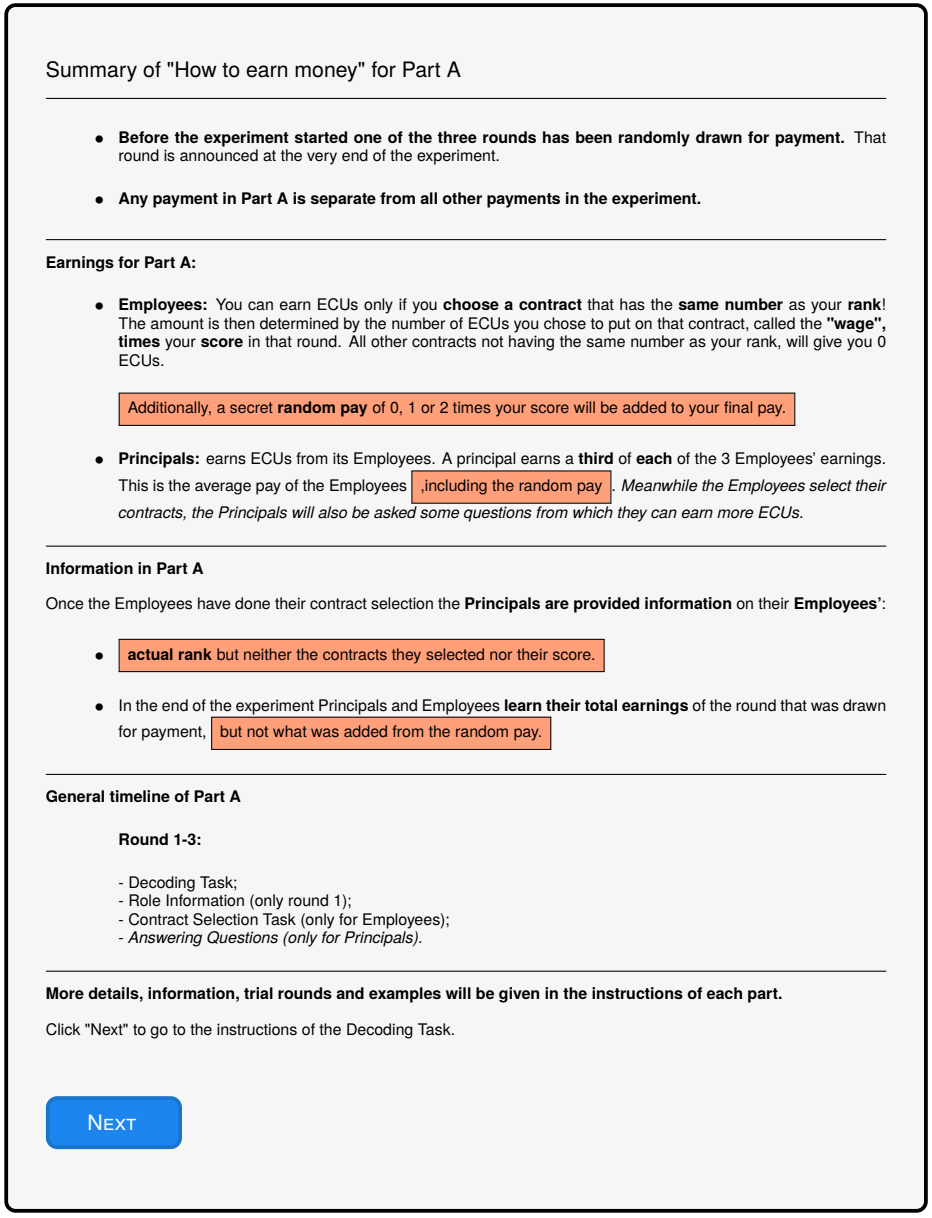
The score, the rank and Selecting a Contract:

- **Score:** In the Decoding Task, each correctly solved problem gives one "point" and will be added up to the Employee's **score** of that round.
- **Rank:** All the 18 Employees' **scores** will automatically be **ranked** such that **rank 1** is assigned to the Employee with the **highest score** and so on until **rank 18** is assigned to the Employee with the **lowest score.** Note that in case there is a score tie between two or more Employees, everyone gets the same higher rank.
- **Contract Selection:** In the Contract Selection, **each Employee chooses** how to **distribute** a total of **[19] ECUs** among 18 available contracts named **Contract 1-18**.

Click "Next" to go to the overview of **How to earn money in Part A.**

NEXT

Figure F.15: page 5



Summary of "How to earn money" for Part A

- **Before the experiment started one of the three rounds has been randomly drawn for payment.** That round is announced at the very end of the experiment.
- **Any payment in Part A is separate from all other payments in the experiment.**

Earnings for Part A:

- **Employees:** You can earn ECUs only if you **choose a contract** that has the **same number** as your **rank**! The amount is then determined by the number of ECUs you chose to put on that contract, called the **"wage", times** your **score** in that round. All other contracts not having the same number as your rank, will give you 0 ECUs.

  Additionally, a secret **random pay** of 0, 1 or 2 times your score will be added to your final pay.

- **Principals:** earns ECUs from its Employees. A principal earns a **third** of **each** of the 3 Employees' earnings. This is the average pay of the Employees ,including the random pay . *Meanwhile the Employees select their contracts, the Principals will also be asked some questions from which they can earn more ECUs.*

Information in Part A

Once the Employees have done their contract selection the **Principals are provided information** on their **Employees'**:

- **actual rank** but neither the contracts they selected nor their score.
- In the end of the experiment Principals and Employees **learn their total earnings** of the round that was drawn for payment, but not what was added from the random pay.

General timeline of Part A

Round 1-3:

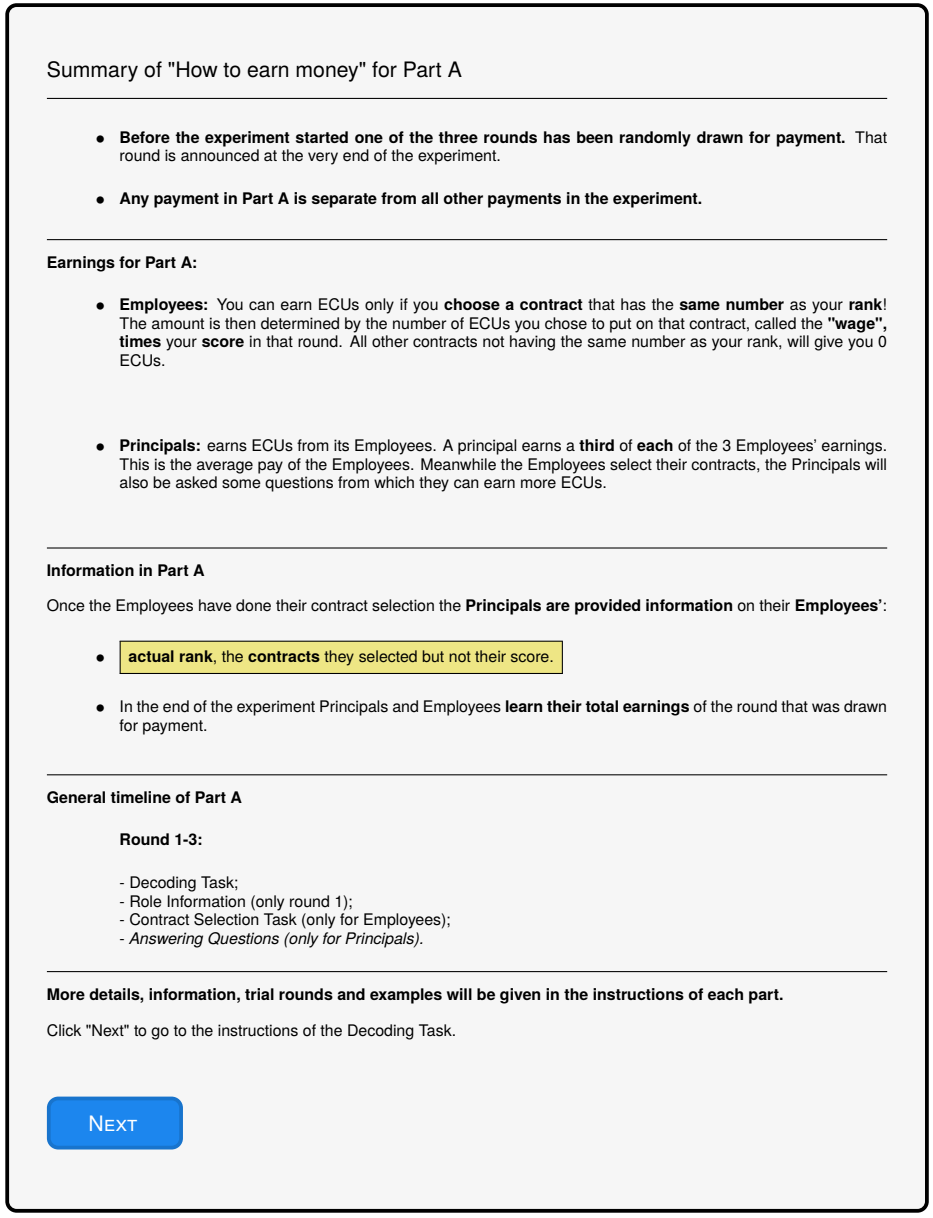
- Decoding Task;
- Role Information (only round 1);
- Contract Selection Task (only for Employees);
- *Answering Questions (only for Principals).*

More details, information, trial rounds and examples will be given in the instructions of each part.

Click "Next" to go to the instructions of the Decoding Task.

NEXT

(a.) Control Treatment



Summary of "How to earn money" for Part A

- **Before the experiment started one of the three rounds has been randomly drawn for payment.** That round is announced at the very end of the experiment.
- **Any payment in Part A is separate from all other payments in the experiment.**

Earnings for Part A:

- **Employees:** You can earn ECUs only if you **choose a contract** that has the **same number** as your **rank**! The amount is then determined by the number of ECUs you chose to put on that contract, called the **"wage", times** your **score** in that round. All other contracts not having the same number as your rank, will give you 0 ECUs.

- **Principals:** earns ECUs from its Employees. A principal earns a **third** of **each** of the 3 Employees' earnings. This is the average pay of the Employees. Meanwhile the Employees select their contracts, the Principals will also be asked some questions from which they can earn more ECUs.

Information in Part A

Once the Employees have done their contract selection the **Principals are provided information** on their **Employees'**:

- **actual rank**, the **contracts** they selected but not their score.
- In the end of the experiment Principals and Employees **learn their total earnings** of the round that was drawn for payment.

General timeline of Part A

Round 1-3:

- Decoding Task;
- Role Information (only round 1);
- Contract Selection Task (only for Employees);
- *Answering Questions (only for Principals).*

More details, information, trial rounds and examples will be given in the instructions of each part.

Click "Next" to go to the instructions of the Decoding Task.

NEXT

(b.) Public Treatment

Figure F.16: page 6

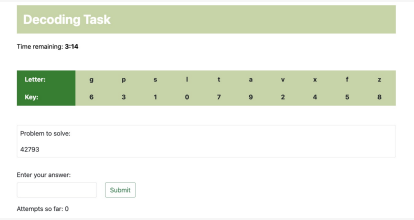
Decoding Task Instructions

The time period for a **Decoding Task** is **4 minutes**. Only a correct and submitted problem gives one point and incorrect answers do not give minus points. New problems appear automatically on the screen, as soon as an answer is submitted.

Example Screen

Below we show an example of a problem. You can see four things:

1) On the top row, the letters where you will search for your answer;
2) On the second row, the decoding key showing which number to match with which letter;
3) In the white box, the "Problem to solve" - the number series to translate to letters and;
4) The answer box where the letter combination should be entered before submitting the answer.



In the example, the problem to solve is to decode 42793. Following the decoding key, the 4 corresponds to the letter x, the 2 to the v, 7 to t, 9 to a and the 3 to the p. To get one point, xvtap needs to be entered in the answer-box and submitted.
At the top, a clock will indicate the remaining time and at the bottom, "Attempts so far" shows how many problems you have tried so far (both correct and incorrect).

Click "Next" to go to the **trial round** of the Decoding Task.

NEXT

Figure F.17: page 7

Trial round

Now you can **try** the Decoding Task for 2 minutes!

During this time period it does not matter if you are right or wrong: it will not affect any of your final earnings. In the trial round you can also see if your last answer was right or wrong, which will not be the case in the real rounds.

As soon as you click "Start Trial" the 2 minutes will start.

START TRIAL

Figure F.18: page 8

Figure F.19: page 9

Figure F.20: page 10

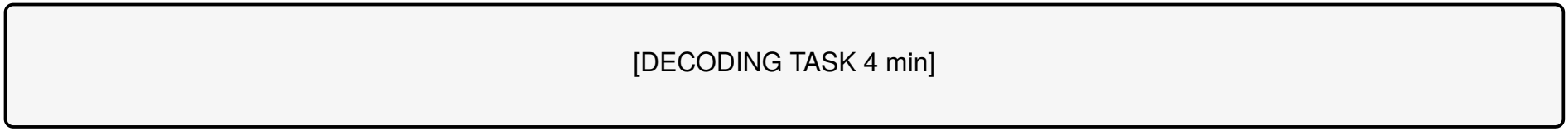


Figure F.21: page 11

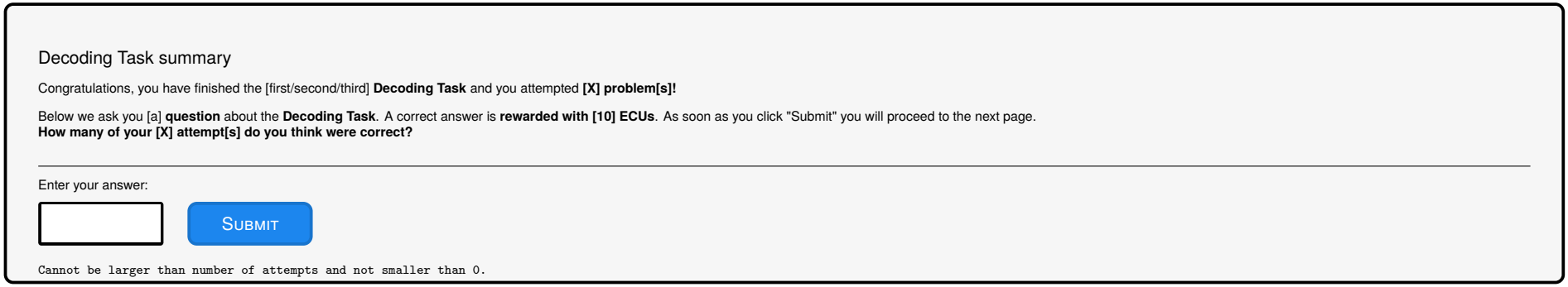


Figure F.22: page 12

Figure F.23: page 13

Contract Selection Instructions

Below we will explain how to select working contract(s) that will determine your earnings. All the 18 Employees repeat this task in all the three rounds.

Score rank: In each round, the Decoding Task scores will automatically be ranked, but are not revealed. The Employee with the **highest score** receives **rank 1** while the second highest scoring Employee receives **rank 2** etc. until the lowest scoring Employee may get **rank 18**. For example, a rank 6 means that 5 other Employees had a higher score. Note that the higher the score, the lower the rank!

Score ties: If two (or more) Employees have the **same score** they receive the **same rank** since there are as many Employees with **higher scores** compared to them. In the example table below you can also see that if there is a tie among two Employees with the very lowest score, the highest received rank is 17 not 18.

Rank example table

Employee	**Decoding Task score**	Rank
E16	100	1
E2	50	2
E1	50	2
E4	10	4
.	.	
.	.	
E3	4	16
E9	1	17
E13	1	17

The Different Contracts: You will have 18 contracts to choose from, numbered from 1-18 (Contract 1; Contract 2; . . . ; Contract 18). You can choose to select only one or all of the eighteen contracts. But, there is only one contract in each round that can give an Employee any earnings - **a contract only pays-off if its number is the same as the rank of that Employee!**

Distributing ECUs: In each round you will also choose how much ECUs to put on each of your chosen contracts. You have **19 ECUs to allocate** to your selected contact/s and there are **two requirements** that need to be fulfilled.

- **All the 19 ECUs must be used.**
- **One contract must get at least 1 ECU more compared to another.**

The "Most Preferred Contract": One Contract will be considered and called the **Most Preferred Contract**. It is the contract that an Employee has chosen to put the most ECUs on – the contract that fulfils the second requirement above.

ECU earnings of a Contract: The ECUs you allocate to a contract is called your **wage** (the pay per point). If a selected contract has the same number as your rank in that round, you will earn that **contract's wage times your score,** if that round is selected for payment.

Contract Selection Trial

Below you can try out different **Contract Selections** for a maximum of 4 minutes, or you can skip this by clicking "Next". The time starts counting down as soon as you make any choice.

The contract selection is done in two steps:

In **Step 1**, you select which of the 18 Contracts you would like and click "Select Contract(s)".
In **Step 2**, you will allocate the 19 ECUs (according to the two requirements) to your selected contracts of Step 1.
A number on the top right is showing how many ECUs you have left to decide about. If this figure becomes red and negative, you have allocated too many ECUs to the contract(s).

As soon as you would like to finalize your selection and allocation of ECUs, click "Submit Selection" and a summary of your choices will appear. If you would like to change your selected contracts of Step 1, click "Deselect Contract(s)".

In this trial, but not in the real rounds, you are able to select contracts as many times as you want during 4 trial minutes by clicking "New Trial Selection".

Select Contract(s) Trial:

Remaining trial time: 3:59

Contract 1 ◯	Contract 2 ◯	Contract 3 ◯	Contract 4 ◯	Contract 5 ◯	Contract 6 ◯
Contract 7 ◯	Contract 8 ◯	Contract 9 ◯	Contract 10 ◯	Contract 11 ◯	Contract 12 ◯
Contract 13 ◯	Contract 14 ◯	Contract 15 ◯	Contract 16 ◯	Contract 17 ◯	Contract 18 ◯

SELECT CONTRACT(S)

Unfold once contracts are selected in Step 1.

Select Contract(s) Trial:

Remaining trial time: 3:23

- Insert a number between 1 and 19 for all of your Selected Contract(s).
- Remember to use all your 19 ECUs.
- Allocate (at least) one more ECU to one of the contract, compared to any other.

DESELECT CONTRACT(S)

ECUs left to allocate: **19 ECUs**

☐ **Contract 3:** You will be paid [X] ECUs/point **if your actual rank was 3**;

☐ **Contract 4:** You will be paid [X] ECUs/point **if your actual rank was 4**;

☐ **Contract 5:** You will be paid [X] ECUs/point **if your actual rank was 5**;

☐ **Contract 6:** You will be paid [X] ECUs/point **if your actual rank was 6**;

SUBMIT SELECTION

Summary: You selected [X] Contract(s) and your selection is valid. Your **Most Preferred Contract** was **Contract [Y]**, to which you allocated [YY] ECUs.

If you would like to try a new Contract Selection click “New Trial Selection".

NEW TRIAL SELECTION

Click "Next" to go to the last information about Selecting a Contract.

NEXT

Figure F.24: page 14

Summary information

When the Decoding Task and the Contract Selection of a round is finished the **Employee** will only get to know about their own **score** and knows their Selected Contract(s), while the **Principal** only get information about the **actual rank** of all its 3 Employees.

Below we show how this will look like for both the Employee and the Principal in a round 3 summary, which also includes information on the previous two rounds.

Example of an Employee's screen in Round 3

This is the summary of last round, including how you selected contracts.

Your score last round was 14.

ECU/Contract ("wage"[1])	Selected Contract(s)
10 ECU	8
5 ECU	7
3 ECU	5
1 ECU	6

Most Preferred Contract

In round 2: your **Most Preferred Contract** was 10.
In round 1: your **Most Preferred Contract** was 7.
*1) the **wage** only pays-off when the Rank and the Contract number equals.*

After the summary Employees are then asked to send some of the information to their principal. The **Principals** get a summary of the 3 Employees with the information below.

Example of a Principal's screen in round 3

Summary of round 3:
This is the summary of last round with the actual **ranks** of your 3 employees as well as a repetition of their round 2 and 1 rankings.

Current Round 3	*Round 2*	*Round 1*
Rank (actual)	Rank (actual)	Rank (actual)
Emp. 1	8	9	10
Emp. 2	3	2	3
Emp. 3	7	6	6

At the very end of the experiment, Employees and Principals will also learn their total earnings of the round that was randomly selected for payment of Part A. Next, we will ask you 5 questions about earnings and the contract selection, to make sure everybody understands. Click "Next" to go to the comprehension test.

NEXT

(a.) Control

Summary information

When the Decoding Task and the Contract Selection of a round is finished the **Employee** will only get to know about their own **score** and knows their Selected Contract(s), while the **Principal** get information about the **actual rank, the** **Most Preferred Contract** (the contract an Employee assigned the most ECUs), the **Difference** (between the actual rank and the Most Preferred Contract) and the **Direction** of the Difference (if the Difference was an under-, accurate or over-estimation) of all its 3 Employees.

Below we show how this will look like for both the Employee and the Principal in a round 3 summary, which also includes information on the previous two rounds.

Example of an Employee's screen in Round 3

This is the summary of last round, including how you selected contracts.

Your score last round was 14.

ECU/Contract ("wage"[1])	Selected Contract(s)
10 ECU	8
5 ECU	7
3 ECU	5
1 ECU	6

Most Preferred Contract

In round 2: your **Most Preferred Contract** was 10.
In round 1: your **Most Preferred Contract** was 7.
*1) the **wage** only pays-off when the Rank and the Contract number equals.*

After the summary Employees are then asked to send some of the information to their principal. The **Principals** get a summary of the 3 Employees with the information below.

Example of a Principal's screen in round 3

Summary of round 3:
This is the summary of last round with the actual **ranks** of your 3 employees as well as a repetition of their round 2 and 1 rankings.

In the table you can also see the information they have sent you, on the **Most Preferred Contract** (the contract given the most ECUs); as well as the **Difference** (between actual rank and the Most Preferred Contract) and the **Direction** of that difference with $+$, $-$ and ± 0 indications for an over-, under- and an accurate- estimation, respectively).

Current Round 3				
	Rank (actual)	Contract (preferred)	Diff. (Rank - MPC)	Direction (of Diff.)
Emp. 1	8	5	$+3$	Overestimation
Emp. 2	3	2	$+1$	Overestimation
Emp. 3	7	9	-2	Underestimation

Round 2				
	Rank	MPC	Diff.	Dir.
Emp. 1	9	14	-5	Underest.
Emp. 2	2	2	± 0	Accurate est.
Emp. 3	6	4	$+2$	Overest.

Round 1			
Rank	MPC	Diff.	Dir.
10	7	$+3$	Overest.
3	3	± 0	Accurate est.
6	8	-2	Underest.

At the very end of the experiment, Employees and Principals will also learn their total earnings of the round that was randomly selected for payment of Part A. Next, we will ask you 5 questions about earnings and the contract selection, to make sure everybody understands. Click "Next" to go to the comprehension test.

NEXT

(b.) Public Treatment

Figure F.25: page 15

Comprehension test

Here is an example of a Decoding Task and a Contract Selection outcome. In the example, there are only 4 instead of 18 Employees, called E1, E2, E3 and E18. You need to answer all the 5 questions correctly to proceed.

Example outcome:

	Score (total points)	**Rank** (actual)	**Selected Contract(s)** (allocated ECUs)	**Wage** (earning/point)	**Earnings** (total pay)
E1	7	3	Contract 6 (10 ECUs); Contract 5 (5 ECUs); Contract 4 (4 ECUs)	0 ECUs	0 ECUs
E2	6	4	Contract 5 (8 ECUs); Contract 4 (6 ECUs); Contract 3 (4 ECUs); Contract 6 (1 ECUs)	? ECUs	? ECUs
E3	9	1	Contract 2 (15 ECUs); Contract 1 (4 ECUs)	4 ECUs	? ECUs
E18	8	2	Contract 3 (9 ECUs); Contract 1 (5 ECUs); Contract 2 (5 ECUs)	5 ECUs	? ECUs

1. **Which is the "Most Preferred Contract" of E1?**

   Contract 3 ◯ Contract 4 ◯ Contract 5 ◯ Contract 6 ◯

2. **Which is the "Most Preferred Contract" of E2?**

   Contract 3 ◯ Contract 4 ◯ Contract 5 ◯ Contract 6 ◯

3. **What is the wage of E2?**

   4 ECUs ◯ 6 ECUs ◯ 8 ECUs ◯ 1 ECU ◯

4. **How much will E3 earn in total (for this example round)?**

   9 ECUs ◯ 135 ECUs ◯ 0 ECUs ◯ 36 ECU◯

5. **Who will earn the most of E3 and E18?**

   E3 will earn 4 ECUs more than E18 ◯

   E18 will earn 4 ECUs more than E3 ◯

   They will earn the same ◯

   None of them will earn any ECUs ◯

Answer all the 5 questions and click "Submit". If an answer is incorrect, it is marked and needs to be changed.

SUBMIT

`Indicate where wrong. Cannot pass without all correct.`

Figure F.26: page 16

Your role

In the beginning of the experiment you were randomly assigned to the role: [**Employee/Principal!**]
Note: You will keep the **same role** and **group** (of 1 Principal and 3 Employees) **for Part A**, but not later in the experiment.

Click "Next" to select Contract(s)

NEXT

Figure F.27: page 17

Contract Selection

In round [1] **you scored [X]** in the Decoding Task. You will now select a contract to be paying you for this work!

Your earnings depend on your choices! Likewise, will your Principal earn a third of what you earn plus a third of the other two Employees' payment that your Principal has been matched with.

Remember that the Contract only pays-off **if its number is the same as your actual rank!**

You will first have 18 contracts to select from and then you will have 19 ECUs to distribute over these selected Contract(s).

Information: After you have selected your contract(s) we will ask you to submit information about your actual **Rank** to your Principal.

You will get a summary reminder of your **selected Contract(s)** and your **score** as described before.

Select your Contract(s) below in two steps:

[SELECTS CONTRACT(S) AS OUTLINED IN Figure F.23]

No time-restriction

SUBMIT SELECTION

NEXT

(a.) Control

Contract Selection

In round [1] **you scored [X]** in the Decoding Task. You will now select a contract to be paying you for this work!

Your earnings depend on your choices! Likewise, will your Principal earn a third of what you earn plus a third of the other two Employees' payment that your Principal has been matched with.

Remember that the Contract only pays-off **if its number is the same as your actual rank!**

You will first have 18 contracts to select from and then you will have 19 ECUs to distribute over these selected Contract(s).

Information: After you have selected your contract(s) we will ask you to submit information about your actual **Rank** and your selected **Most Preferred Contract** as well as the **Difference** (Rank-MPC) between those and the **Direction** of the difference to your Principal.

You will get a summary reminder of your **selected Contract(s)** and your **score** as described before.

Note that you will not know your actual **Rank** and neither the **Difference** nor the **Direction** of the difference, when you submit it to your Principal.

Select your Contract(s) below in two steps:

[SELECTS CONTRACT(S) AS OUTLINED IN Figure F.23]

No time-restriction

SUBMIT SELECTION

NEXT

(b.) Public Treatment

Figure F.28: page 18

[SUMMARY OF ROUND 1/2/3 as outlined in upper part of Figure F.24a. and F.24b.]

Figure F.29: page 19

Submitting information

Information: Below there is **information** that we ask you to **send to your Principal**, together with the information that you have finished round 2.

The principal will get information about your **actual rank**. To finish click "Submit information to my Principal".

Submission to my Principal

I have now finished round 2 and I am submitting information about my **rank**.

Submit information to my Principal

(a.) Control

Submitting information

Information: Below there is **information** that we ask you to **send to your Principal**, together with the information that you have finished round 2.

The principal will get information about your **actual rank** and we ask you to fill in your "Most Preferred Contract" of this round in the box. To finish click "Submit information to my Principal".

Submission to my Principal

I have now finished round 2 and I am submitting information about my **rank**.

:_____ as my **Most Preferred Contract**.

Submit information to my Principal

(b.) Public Treatment

Figure F.30: page 20

End of Round

You have now finished round [1/2/3] and you will now be waiting for the others to finish this round as well.

Click "Next" to proceed.

Next

[After Round 3 – Other Exp. parts.]

Figure F.31: page 21

Welcome to PART B!

Click "Next" to start this part of the experiment

NEXT

Figure F.32: page 22

Instructions to Part B:

This is the last part of the experiment and consists of two short stages, a lottery choice and two short surveys. You can win money in the lottery.

Click "Next" to go to the lottery.

NEXT

Figure F.33: page 23

Lottery Choice

The Lotteries: You are now given the option to choose between two lotteries: one with a lower gain of 300 ECU, but with a higher probability of winning or one with a higher gain of 700 ECU, but with a lower probability to win.

You can choose only one of these lotteries. In the first lottery, Lottery A, you must choose any number between 1-50 and in the second, Lottery B, you must choose any number between 1-100.

In the end of the experiment we will randomly draw the winning number for both Lottery A and B, and announce the winning number on everybody's screen. You will be shown individually whether you won or lost.

Pick a number from one of the lotteries that you want to play in. Once you click "Submit", you cannot change your mind and you will be redirected to the survey. Please choose a number from one of the lotteries by clicking that number:

Lottery A

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31	32	33	34	35
36	37	38	39	40
41	42	43	44	45
46	47	48	49	50

Lottery B

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31	32	33	34	35
36	37	38	39	40
41	42	43	44	45
46	47	48	49	50
51	52	53	54	55
56	57	58	59	60
61	62	63	64	65
66	67	68	69	70
71	72	73	74	75
76	77	78	79	80
81	82	83	84	85
86	87	88	89	90
91	92	93	94	95
96	97	98	99	100

You have chosen to play in Lottery [B] and you have picked number [42]. If this is correct, submit your bet below by clicking "Select".

SELECT

[Short experimental survey.]

Figure F.34: page 24

Short Survey II

You will now be given five questions to answer:

Below we ask you three questions about your views.

We ask you to provide your answers on a scale of how much you agree with the statement. The value 0 means that you do not agree at all, indicated with "Strongly disagree"; and the value 10 means that you completely agree, indicated with "Strongly agree", with the statement.

S1. Vaccines weaken the immune system

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Strongly disagree — Strongly agree

S2. Vaccines can cause the illness they are meant to prevent (e.g., you vaccinate against the flu and that gives you the flu)

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Strongly disagree — Strongly agree

S3. Vaccines could cause serious and permanent side effects

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Strongly disagree — Strongly agree

As soon as you click "Submit" your answers will be registered and you cannot go back and change them.

SUBMIT

Figure F.35: page 25

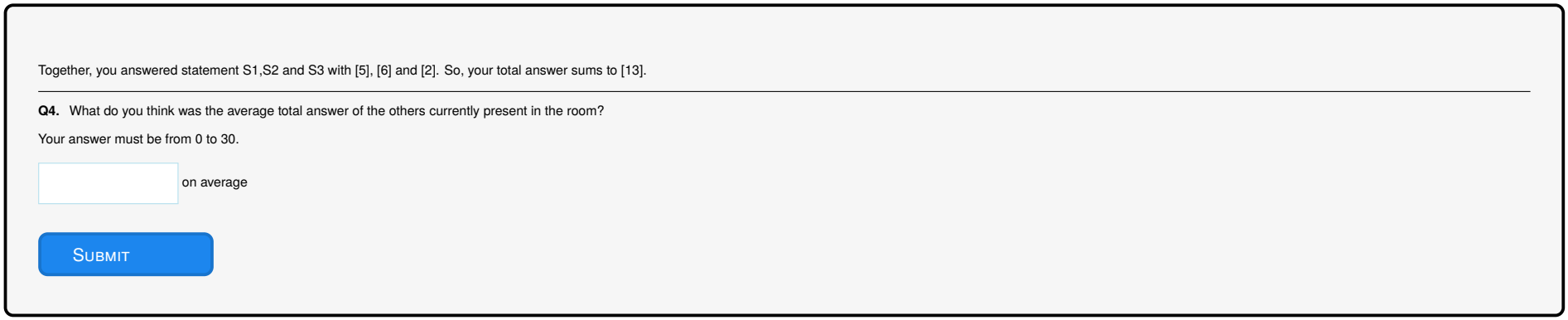


Figure F.36: page 26

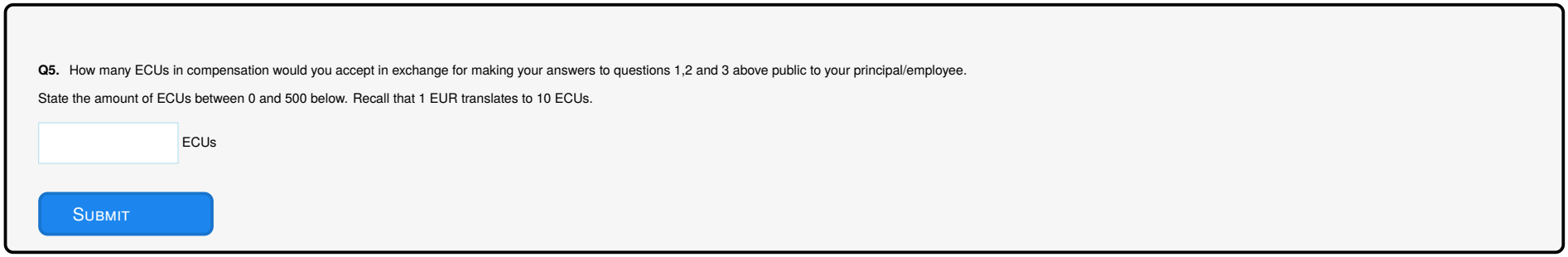


Figure F.37: page 27

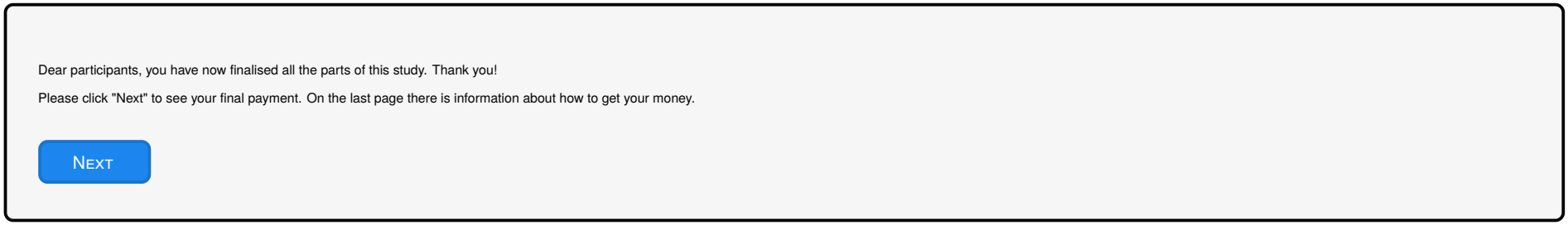
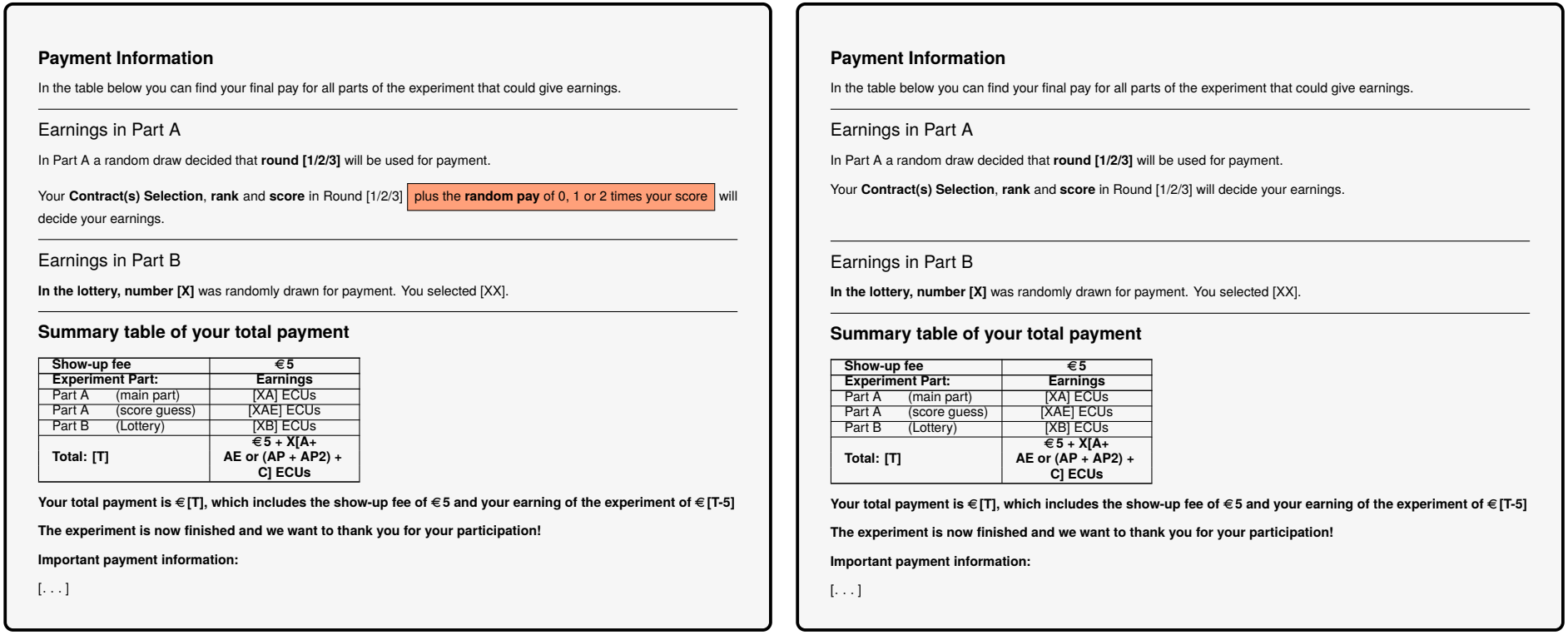


Figure F.38: page 28



Payment Information

In the table below you can find your final pay for all parts of the experiment that could give earnings.

Earnings in Part A

In Part A a random draw decided that **round [1/2/3]** will be used for payment.

Your **Contract(s) Selection**, **rank** and **score** in Round [1/2/3] plus the **random pay** of 0, 1 or 2 times your score will decide your earnings.

Earnings in Part B

In the lottery, number [X] was randomly drawn for payment. You selected [XX].

Summary table of your total payment

Show-up fee		**€ 5**
Experiment Part:		**Earnings**
Part A	(main part)	[XA] ECUs
Part A	(score guess)	[XAE] ECUs
Part B	(Lottery)	[XB] ECUs
Total: [T]		**€ 5 + X[A+ AE or (AP + AP2) + C] ECUs**

Your total payment is € [T], which includes the show-up fee of € 5 and your earning of the experiment of € [T-5]

The experiment is now finished and we want to thank you for your participation!

Important payment information:

[. . .]

(a.) Control

Payment Information

In the table below you can find your final pay for all parts of the experiment that could give earnings.

Earnings in Part A

In Part A a random draw decided that **round [1/2/3]** will be used for payment.

Your **Contract(s) Selection**, **rank** and **score** in Round [1/2/3] will decide your earnings.

Earnings in Part B

In the lottery, number [X] was randomly drawn for payment. You selected [XX].

Summary table of your total payment

Show-up fee		**€ 5**
Experiment Part:		**Earnings**
Part A	(main part)	[XA] ECUs
Part A	(score guess)	[XAE] ECUs
Part B	(Lottery)	[XB] ECUs
Total: [T]		**€ 5 + X[A+ AE or (AP + AP2) + C] ECUs**

Your total payment is € [T], which includes the show-up fee of € 5 and your earning of the experiment of € [T-5]

The experiment is now finished and we want to thank you for your participation!

Important payment information:

[. . .]

(b.) Public Treatment

G Robustness check: alternative misbelief scales

Table G.4 below the correlation between CATEs and binary misbeliefs about vaccines measures, where the latter are constructed excluding values closer to the midpoint. In practice, from the full 11-item Likert scale of agreement, we consider subjects to agree if they selected values equal to or greater than 7 (the corresponding threshold in the main analysis is 6). All the results are qualitatively confirmed and, for the misbelief that vaccines cause illness, the statistical significance is improved from 10% to 5%, and the magnitude of the correlation with the CATE for placement ("overconfidence") increases from -14.8 percentage points to -18.7 percentage points.

Table G.4: Correlation of CATEs and misbeliefs: alternative scales

Agreement with	**Coefficient**	**s.e.**
Correlation with CATE for mrb ("modesty")		
Misbelief: illness causation	0.253**	(0.128)
Misbelief: adverse effects	0.022	(0.150)
Correlation with CATE for placement ("confidence bias")		
Misbelief: illness causation	-0.187**	(0.070)
Misbelief: adverse effects	-0.058	(0.083)
N	112	

NOTES: Correlations are estimated by bivariate OLS models between the CATEs (individual causal effects of self-consciousness elicitation that embed covariates for their estimation) and an alternative binary measure of agreement with vaccine misbeliefs (coding agreement for values equal to or above 7 on an 11-item Likert scale). *p<0.1; **p<0.05; ***p<0.01.