The Inference Cost of Policies[*]

Gonzalo Arrieta[†] Maxim Bakhtin[‡]

October 12, 2025

Abstract

Screening decisions can be crucial in various situations, such as hiring workers, selling insurance, or designing policies. If a policy encourages uniform behavior, different people behave similarly, which makes learning about them from their behavior impossible. Choosing whether to screen can be difficult because it requires recognizing and trading off the benefits of information with the costs of getting it. We investigate whether people solve this trade-off optimally and what causes their mistakes. We design an online experiment that simulates a hiring scenario with an initial trial task. Participants make two decisions: first, they select a trial task, which can reveal candidates' quality at a small cost, and then choose which candidate to hire. We show that most participants choose the suboptimal task that does not reveal the candidates' quality, and this mistake persists even with experience and feedback. We test for the mechanisms and show that insufficient screening is driven by failures to anticipate inference and plan the full strategy.

[*] We thank B. Douglas Bernheim, Lukas Bolte, John Conlon, Stanton Hudja, Muriel Niederle, Kirby Nielsen, Alvin Roth, Jason Somerville, and seminar participants at BABEEW, BEESCuitS, ESA, IZA, and Stanford University for their helpful comments. This study is covered under Stanford's IRB Protocol 44866 and was registered on the AEA RCT registry under ID AEARCTR-0012230 titled "The inference cost of interventions."

[†] gonzalo.arrieta@econ.uzh.ch; Department of Economics, University of Zurich

[‡] Analysis Group, Inc.

I. Introduction

Screening is important and prevalent in many economic environments. When hiring a new employee, selling insurance, or buying a good of unknown quality, individuals and organizations take costly steps to learn more about the other party, like doing lengthy interviews with candidates or offering them multiple contracts at different prices. In many other situations, however, people do not screen and act with limited information.

The screening decision involves a trade-off. Effective screening (i.e., policies that lead to different people taking different actions) may be costly, but it reveals valuable information. Assessing the extent to which the informational gains outweigh the costs requires comparison of all relevant contingencies, which recent literature shows people have difficulties in doing (Niederle and Vespa, 2023). To illustrate, consider an employer who hires a new employee: Should they assign difficult or easy tasks to the employee? Difficult tasks, if failed, may come at a greater cost for the employer, but they allow the employer to gauge the new hire's abilities. Similarly, this trade-off in screening is essential for policy choices. A stringent policy may force everyone to behave the same way, which impedes drawing inferences about them from their behavior. For example, if prisons strictly enforce stringent rules, every inmate will follow them. This policy may reduce violence within prisons, but it also destroys information about inmates' character and their willingness to follow the law. These signals can be helpful in early parole decisions, but they are lost if every person behaves the same way. Failures of contingent thinking can, then, make it hard to strike the optimum balance between the forces behind this trade-off, leading to suboptimal policies.

In this paper, we investigate the role of failures of contingent thinking in optimal screening. To illustrate how contingent thinking complicates the screening trade-off, let us expand on the example of an employer deciding whether to assign a difficult or an easy task to a new employee. On the one hand, assigning a difficult task will reveal information about the employee's abilities, which is helpful for future promotion decisions. On the other hand, there is a higher chance that the employee will fail the task, imposing costs on the employer. In addition to this trade-off, the employer may avoid assigning difficult tasks for reasons unrelated to contingent thinking. First, the employer may heavily discount the future benefits of information or be time inconsistent. Second, the employer may be risk-averse and avoid the environmental uncertainty inherent in assigning a difficult task to a new employee (e.g., uncertainty coming from the nature of the task). Third, strategic considerations, such as uncertainty about the employee's response, may also prevent the employer from assigning a difficult task. These factors make it difficult to empirically identify whether people screen optimally.

We hypothesize and experimentally test that people screen too little, even without these confounding factors. Our hypothesis stems from the intuition that, while the screening *costs* are usually immediate and evident when making a screening decision, the *benefits* of the extra information are only realized in subsequent decisions and are harder to recognize (Niederle and Vespa, 2023). Individuals who do not consider subsequent decisions or do not expect to learn useful information are likely not to screen (enough).

We test our hypotheses using an online experiment. It allows us to create a controlled environment to identify participants' choices as mistakes, which is impossible in an observational study without strong additional assumptions. The ideal setting needs an *initial choice* that reveals information at an implicit cost—the decision to screen or not to screen—followed by a *second choice* that makes the information valuable. Moreover, to argue that screening is optimal, revealing information at the initial stage must be optimal. We design a single-agent, single-period decision problem without uncertainty that satisfies these criteria. Our design ensures we can interpret participants' choices as mistakes rather than preferences.

The experiment mimics a hiring problem with an initial trial task. Specifically, participants see two computers and need to hire one after observing their performance on a trial task. One of the computers is *Good*, and the other is *Bad*, but the participant does not observe which computer is which—the computers' type is the information that the trial task can reveal. The participants make two choices: one in part 1 and another in part 2:

Part 1: The participants choose one of two trial tasks for the computers to complete: a *Screening* task or a *Pooling* task. This choice is our main elicitation of interest. On a Screening task, the Good computer generates a high payoff for the participant, while the Bad computer generates a low payoff. On a Pooling task, both the Good and the Bad computers generate a high payoff for the participant. Thus, the Screening task gives a lower part 1 monetary payoff than the Pooling task, but it reveals the computers' quality, while the Pooling task does not.

Part 2: the participants choose one computer to hire. They receive a higher payoff if they hire the Good computer and a lower payoff if they hire the Bad computer. Those participants who in part 1 chose the Screening task know which computer is Good, but those who chose the Pooling task do not.

This design makes information about the quality of computers valuable. The trade-off between a lower payoff from the Screening task in part 1 and a higher payoff from the informed hiring choice in part 2 determines whether getting the information about the computers' quality is optimal.

The experiment consists of *ten rounds*. We make choosing the Screening task optimal by making the stakes in the part 2 hiring choice much larger than those in the part 1 trial task. Considering parts 1 and 2 jointly, the Screening task ensures a guaranteed high payoff, while the Pooling task induces a lottery with a substantially smaller expected payoff. Thus, participants should prefer the Screening task unless they are extremely risk-loving.[1]

We identify suboptimal screening using two treatments. In the *Baseline* treatment, participants start with the trial task in part 1 and continue to the hiring task in part 2, as described above. We use this treatment to estimate the share of mistakes. However, some mistakes in the *Baseline* treatment may come from the experimental noise (e.g., participants' inattentiveness or trembling hand errors).[2] To estimate the amount of noise, we

[1] Table I lists the complete set of parameters used in all rounds.

[2] We use comprehension checks to verify that the participants understand the setup, as the experimental

run a *Strategy Method* control treatment. This treatment helps participants as much as possible in making their decisions while maintaining the same experiment structure. In this treatment, participants solve the problem backward: they first make the part 2 hiring choices *conditional* on the two possible part 1 trial tasks—the Screening task and the Pooling task—and then choose a trial task in part 1. We also help participants (i) with making the inference about which computer is of which quality based on the information available to them, and (ii) with aggregating all payoff consequences of each task choice. We attribute mistakes in the *Strategy Method* treatment to noise and use it as a benchmark for the amount of experimental noise in the *Baseline* treatment. This imposes a higher bar in testing our hypothesis: for us to conclude that individuals screen suboptimally, it is not sufficient that the rate of mistakes in the *Baseline* treatment is positive, but it must be significantly higher than in the *Strategy Method* treatment.

We run the experiment online with 982 Prolific participants and find evidence that people screen suboptimally. In round 1, 68% of participants make the mistake of choosing the Pooling task in the *Baseline* treatment. In contrast, the mistake rate in the *Strategy Method* treatment—which measures experimental noise—is only 18%. Do feedback and experience help reduce mistakes? Yes, but only partially. The average mistake rate across rounds 2–10 is 34% in the *Baseline* treatment and 21% in the *Strategy Method* treatment. Both differences across treatments are statistically significant. These results show that insufficient screening is prevalent and does not fully disappear with experience.

We investigate two driving mechanisms of insufficient screening: Failure to Anticipate Inference and Failure to Plan.

Failure to Anticipate Inference: We test whether, when choosing the trial task in our experiment, participants fail to anticipate that they will be able to infer the computers' quality. We test this Failure to Anticipate Inference using the *Automatic Inference* treatment. This treatment is identical to the *Baseline* treatment, but participants automatically re-

instructions in Appendix section I.B show. Section IV discusses this further.

ceive explicit information about which computer is Good and which is Bad if they choose the Screening task. Participants know about this at the moment of choosing the task in part 1. This intervention eliminates the need to make inferences (and to anticipate the possibility of an inference) and thus serves as a diagnostic test for the Failure to Anticipate Inference as a mechanism driving mistake rates.

Failure to Plan: We test whether participants fail to plan their full strategy. The optimal choice of the part 1 trial task requires participants to plan their strategy for the subsequent part 2 hiring choice. The Failure to Plan is a mechanism that can drive mistakes, and we test it using the *Plan* treatment. In this treatment, in contrast to *Baseline*, participants *simultaneously* make both the part 1 trial task choice and the part 2 hiring choice. This treatment is intended to force participants to think ahead and consider the entire strategy, thus gauging the Failure to Plan mechanism.

Results of the *Automatic Inference* and the *Plan* treatments show evidence for the proposed mechanisms.[3] In the *Automatic Inference* treatment, the mistake rate is much lower than in *Baseline*—44% in round 1 and 22% on average in rounds 2-10. Similarly, in the *Plan* treatment, the mistake rate is 36% in round 1 and 19% in rounds 2-10. Each mechanism accounts for about half of the round 1 mistakes in *Baseline*. In rounds 2-10, the mistake rates in both mechanism treatments are indistinguishable from the *Strategy Method* treatment. These results suggest that the interaction of the two mechanisms prevents complete learning. Even a partial intervention that eliminates one of the mechanisms could be sufficient to help individuals learn the optimal strategy.

The results are robust to an array of potential concerns. First, we rule out risk-loving as a possible explanation. We elicit participants' preferences over the pairs of induced lotteries corresponding to the Pooling (risky lottery) and Screening (guaranteed payoff) tasks.

[3] While we design each treatment to tackle one mechanism, we cannot rule out the possibility that either treatment allows participants to correct both mechanisms (for example, because the interaction of both mechanisms made it too complex for them to realize, but the elimination of one allows for a full correction). Ultimately, both mechanisms are a consequence of the inherent complexity of the task at hand, and we show evidence that both treatments help participants think through the problem.

For each parameterization, at least 97% of participants choose the guaranteed payoff corresponding to the Screening task, which rules out risk-loving as a possible driver of the results. Second, the results are not driven by confusion or misunderstanding of the instructions. Most participants make few errors in understanding questions and attention checks, and the results remain the same on a subsample of participants who make zero errors. Third, the mistakes are not driven by suboptimal information use. If participants do not know how to optimally use the information in part 2, this could justify choosing the Pooling task in part 1. Hiring mistakes in part 2 are rare, suggesting participants can use the information optimally, and the results in round 1 are unchanged for the subsample of participants who choose the Screening task at least once and always use the revealed information correctly. Furthermore, our design ensures no scope for time preferences to rationalize the Pooling task choice, because all payments occur at the end of the experiment. Our design also leaves no room for strategic reasoning concerns, since the computers' behavior is deterministically pre-determined. Lastly, the results are also robust to controlling for demographic characteristics and education.

We relate to the literature on failures of contingent thinking, which shows that people make mistakes when they need to evaluate multiple hypothetical scenarios (Esponda and Vespa, 2014; Martínez-Marquina et al., 2019). Niederle and Vespa, 2023 focus on a cognitive limitation by which individuals have problems holding more than one state in their mind. However, many of those same individuals are able to focus on two states when states are realized rather than hypothetical. We provide two conceptual and one applied contribution. First, our results show that the difficulty of contingent thinking applies not only to exogenous hypothetical events (like those in Esponda and Vespa, 2024) but also to events arising from an individual's own choices. Chakraborty and Kendall, 2022 are closest to us in providing empirical evidence about how boundedly-rational agents fail to perform backwards induction in settings in which they need to understand—and respond to—*their own* decisions at hypothetical future events. They experimentally demonstrate that participants evaluating uncertain artificial investments behave as if they expected to take dominated actions in the future. In our setting, participants do not exactly mispredict their future actions, but rather mispredict how their current actions can inform their future ones. Moreover, unlike seminal experiments in this area (like those discussed in Niederle and Vespa, 2023), in our setting, the resolution of uncertainty is directly tied to the decision-maker's choice. Our results suggest that tying the relevant contingencies to the decision-maker's choice in this way does not fully remove the difficulties they face in placing themselves in such contingencies. Second, we show that having realized instead of hypothetical events does not always get rid of the problem: The *Good* and the *Bad* computers in our experiment are realized, resembling the deterministic treatments that help solve failures of contingent thinking in Martínez-Marquina et al., 2019, yet participants in our setting make substantial mistakes.[4] Our *Strategy Method*, *Automatic Inference* and *Plan* treatments suggest participants have difficulties in thinking about the contingencies in part 2, which are deterministically predetermined by their own part 1 choices, thereby speaking to the prevalence of this type of cognitive mistake.[5] Third, we illustrate the costs of failures of contingent thinking in the context of screening for hiring. This connection sets the stage for studies that can mimic our design in applied environments, contributing to a limited literature studying these cognitive biases in the field, and specifically focusing on screening, which is important and prevalent in many economic environments (while Tergiman, 2024 study a college admissions problem, and Bhargava et al., 2017 study health insurance choices, we know of no other study yet that focuses on failures of contingent thinking applied to screening for hiring).

We are closely related to the experimental literature on learning and bandit problems. This literature studies the exploration-exploitation trade-off—the choice between actions,

[4] While the existence of a *Good* and a *Bad* computer is deterministic, their identity in part 2 is not. It is possible that this is the contingency that participants struggle with, emphasizing the extent to which failures of contingent thinking can manifest.

[5] Our setting adds to the example from Ellsberg, 1961, as discussed by Niederle and Vespa, 2023, in which failures of contingent reasoning happen in situations in which agents do not have to construct any payoff mapping from actions and states, because payoffs for each action-state pair are already provided to participants (in contrast with the Committee Voting and Acquiring-a-Company problems, also discussed by Niederle and Vespa, 2023.

some of which have unknown distributions of payoffs. Most bandit literature is theoretical and concerned with finding the right balance between exploiting actions with known rewards and exploring new actions with uncertain rewards (Bergemann and Välimäki, 2018; see Slivkins et al., 2019 for a modern introduction). Empirically, there is no consensus on whether people under-explore or over-explore. Some papers find evidence of under-exploration (Anderson, 2012; Banovetz, 2020; Hudja and Woods, 2024), while other do not (Kwon, 2020; Hoelzemann and Klein, 2021). We contribute by providing further evidence for under-exploration in a relatively more naturalistic framework.[6] Our second contribution to this literature is identifying mechanisms that contribute to under-exploration — Failure to Anticipate Inference and Failure to Plan. Merlo and Schotter, 1999, 2003 show that people learn less when they receive payoffs for their actions. Our experiment illustrates this result in a straightforward, non-strategic setting. A small payoff from the trial task may distract participants from thinking about the hiring stage and the inference required for it.

We also relate to the behavioral literature on failures of strategic reasoning. People fail to make the right choice in strategic settings, where it requires thinking about others' actions and beliefs (Milgrom and Roberts, 1986; Eyster and Rabin, 2005; Esponda and Pouzo, 2017; Dal Bó et al., 2018; Eyster, 2019; Calford and Cason, 2024). We contribute by showing that people fail to consider inference and plan their strategy even in a *non-strategic* setup. In particular, experimental research has shown that people fail to use backward induction in strategic games (Johnson et al., 2002; Binmore et al., 2002; Levitt et al., 2011; Dufwenberg and Van Essen, 2018). One of the mechanisms we identify in our experiment—Failure to Plan—is a non-strategic analog of failure of backward induction, demonstrating that the difficulty of backward induction extends beyond strategic settings.

The rest of the paper is structured as follows. Section II describes the main experimental design and how we test for the mechanisms driving behavior. Section III summarizes the

[6] We interpret screening as exploration in that it reveals immediately instrumental information at a risk of earning lower payoffs. We interpret pooling as exploitation in that it picks the task that gives certain higher payoffs, with no information revelation. Ultimately, in our setting, information is revealed at the hiring stage in part 2 regardless, but at that point information is not valuable.

main results and the evidence for the mechanisms. Section IV shows the robustness of the results. Section V discusses some implications of the results and concludes.

II. Experimental Design

We design an experiment to test whether people screen enough. Our online experiment mimics a hiring decision with an initial trial task stage, in which participants choose between two options, only one of which reveals valuable information at an implicit cost.

II.A. Setting

The experiment needs to have two crucial features to test our hypothesis. First, information must be *instrumental*, which makes learning valuable. Second, there must be a *screening* decision: the participants need to choose a policy where one option provides valuable information—at some cost—while the other does not. This creates the key trade-off in our hypothesis. Do participants suboptimally choose not to screen?

Our experimental design incorporates the two features that we describe above. We frame the experiment as a hiring problem because it is a naturalistic setting where inference is essential. The participant faces two computers and needs to decide which one to hire. One of the computers is of *Good* quality, and the other is of *Bad* quality. Hiring the Good quality computer is optimal, but the participant does not know each computer's quality. However, the participant can learn the quality by observing the computer's performance in a trial task. The participant makes a policy choice: they choose whether the trial task is *Pooling* or *Screening*. If the task is Screening, the participant can infer the computers' quality at an implicit cost. The inference happens by observing the payoff that the computer produces when solving the task. If the task is Pooling, they cannot infer anything.

We use computer players rather than human participants as the experiment's candidates because this rules out confounders and improves identification. First, if candidates were human participants, working on a task would provide experience with that task. This scope for experience could incentivize participants to choose one trial task over the other, which would confound our results. Second, social preferences could affect the participants' choice of trial task and hiring decision, and the direction of this confounder is ambiguous. Third, we need to ensure consistent performance of candidates across tasks, which is what defines the Good and the Bad candidates. Human candidates' performance is likely to be inconsistent, which would cause the trial task to reveal little information. We avoid all these concerns by replacing human candidates with computers. There are no training, social preferences, or inconsistency concerns with computer candidates.

We run four treatments, described in sections II.D and II.E. In the following subsection, we detail the experiment setup that is common to all treatments.

II.B. Experiment Setup

Each round of the hiring problem consists of two parts: part 1 is the trial task, which is the main decision of interest, and part 2 is the hiring decision. Figure I illustrates the structure of each round. The first part, the trial task, allows the participant to infer which candidate is better. The second part, the hiring decision, makes this information valuable. Specifically, two computers, Good and Bad, generate a payoff for the participant by solving tasks. There are two types of tasks: Pooling and Screening. In the Pooling task, both computers generate the same high payoff, P_H. In the Screening task, the Good computer generates a high payoff, P_H, while the Bad computer generates a low payoff, $P_L < P_H$. Thus, the participants can infer the quality of each computer if they observe the payoffs they generate in the Screening task.[7] Choosing the Screening task provides information, but it comes at a cost. Thus, the participant faces a trade-off between receiving a lower payoff from the trial task ($P_H + P_L < 2 \times P_H$) and learning the computers' quality. This trade-off is

[7] That is, if they observe a payoff of P_L (Screening task), they know that the computer that generated it is Bad. If they observe both computers generating P_H (Pooling task), then they cannot infer the computers' quality.

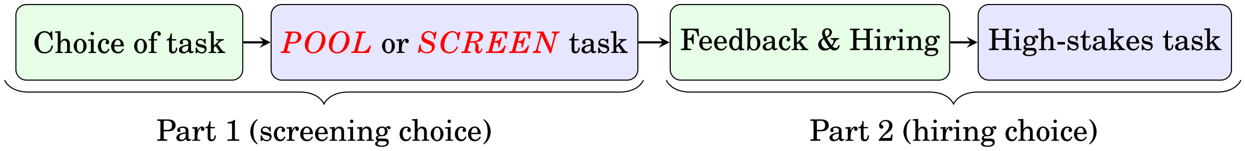


Figure I: Diagram of experiment for Screening-optimal rounds

the crucial element of our design.

After choosing the Pooling or the Screening task in part 1, the participant enters the part 2 hiring decision, in which knowing the computers' quality is valuable because the computers are always hired to solve the Screening task. The stakes in the part 2 Screening task are much higher than in the trial task, which justifies forgoing a part of the payoff in the trial task to learn which computer is Good.[8] More precisely, when screening is optimal, in part 2 the good computer produces P_G, and the bad computer produces $P_B < P_G$ with $P_G > 2 \times P_H - (P_H + P_L)$. Hence, if the participant chooses the Screening task in part 1 and hires the Good computer in part 2, they get $P_H + P_L$ in part 1, and P_G in part 2, with certainty. If the participant instead chooses the Pooling task in part 1, then in part 2 they face a lottery with equally likely payoffs of $2 \times P_H + P_G$ and $2 \times P_H + P_B$. Hence, the Screening task is optimal unless the participant is sufficiently risk-loving.[9]

Participants face ten rounds with varying payoff parameters. Table I summarizes the parameter values in the ten parameterizations. The parameterization in the first and last rounds are the same for all participants, but parameterizations appear in random order in rounds 2–9. Parameterizations 1–6, which includes the first and last rounds, follow the description above, where it is optimal to choose the Screening task.[10] Parameterizations 7–10 make it optimal to choose the Pooling task. While our primary interest is the Screening-optimal rounds, we include Pooling-optimal rounds to prevent participants from

[8] We discuss the optimal choice in the trial task in subsection II.C.

[9] The only change needed for the Pooling task to be optimal is that parameters are such that $P_G < 2 \times P_H - (P_H + P_L)$.

[10] For example, we use parameterization 1 in the first round: In the Pooling task, both computers generate \$0.05. In the Screening task, the Good computer generates \$0.05, while the Bad computer generates \$0.00. In part 2, the Good computer generates \$4.30, while the Bad computer generates only \$0.05.

mechanically choosing the Screening task without regard for the payoffs. Only one randomly picked parameterization counts toward the participant's payoffs. We label the Pooling and the Screening tasks with color names (e.g., Brown and Blue tasks), and vary them every round.[11]

	Part 1			Part 2		
Param. #	$POOL$, both	$SCREEN$, Good	$SCREEN$, Bad	Good	Bad	Optimum
1	0.05	0.05	0.00	4.30	0.05	$SCREEN$
2	0.05	0.05	0.00	4.45	0.10	$SCREEN$
3	0.20	0.20	0.15	4.30	0.10	$SCREEN$
4	0.05	0.05	0.00	4.45	0.10	$SCREEN$
5	0.05	0.05	0.00	4.35	0.10	$SCREEN$
6	0.05	0.05	0.00	4.50	0.10	$SCREEN$
7	2.20	2.20	0.15	0.20	0.20	$POOL$
8	2.10	2.10	0.00	0.05	0.05	$POOL$
9	2.00	2.00	0.00	0.05	0.05	$POOL$
10	2.15	2.15	0.00	0.05	0.05	$POOL$

Table I: Summary of payoff parameters for the ten parameterizations. All values are in USD terms. The parameterization in the first and last rounds are the same for all participants, and randomly drawn from parameterizations 1–6. In rounds 2–9, parameterizations are randomly drawn from all remaining ones.

At the beginning of the experiment, we thoroughly explain its structure to the participants so they have all the information to make optimal choices. We carefully explain what decisions the participants will face, what payoffs they can receive, and what information they will have. Importantly, the participants know that they will observe the performance of each computer before the part 2 hiring stage. We employ rigorous comprehension checks to verify that the participants understand the setup.[12]

[11] This further randomizes the positioning of the tasks in the screen, preventing participants from anchoring to an initial option.

[12] All experimental instructions are in Appendix section I.B.

II.C. What is a mistake?

Parameterizations 1–6: Participant's risk preferences determine whether Pooling is a mistake or not. The two part 1 trial tasks induce two different lotteries, and we show that participants' risk preferences are such that choosing the Pooling task in the Screening-optimal rounds is indeed a mistake. Time preferences do not affect the choice of the trial task because all payments happen at the end of the experiment. Moreover, this is a single-person decision problem, so strategic considerations play no role.

To illustrate the payoff consequences of a Screening-optimal round, consider the first round parameterization (row 1 of Table I), which is the same for everyone: If the participant chooses the Screening task in part 1 and hires the Good computer in part 2, she gets \$4.35 with certainty (\$0.05 + \$0.00 in part 1 and \$4.3 in part 2). If the participant instead chooses the Pooling task in part 1, then in part 2 she faces a lottery with equally likely payoffs of \$4.40 (\$0.05 + \$0.05 in part 1 and \$4.3 in part 2) and \$0.15 (\$0.05 + \$0.05 in part 1 and \$0.05 in part 2).[13] Hence, the Screening task is optimal unless the participant is sufficiently risk-loving.[14]

We diagnose the strength of our assumption that the Screening task is optimal by directly eliciting participants' risk preferences. Assuming that a direct choice over lotteries better captures participants' risk preferences, this exercise supports our claim that choosing the Pooling task is indeed a mistake.[15] We elicit participants' preferences over the lotteries induced in each parameterization by, after the main experiment, having them face six binary menus of lotteries. Each menu has two lotteries induced by one of the six Screening-optimal parameterizations: one lottery is a fixed payoff resulting from choosing the Screening task

[13] In Screening-optimal rounds, the Pooling task reduces the participants' expected payoff by at least \$2 compared to the Screening task, which represents 53% of their average bonus payment.

[14] Assuming CRRA utility, if a person prefers the lottery induced by the Pooling task, they would also prefer a lottery that pays \$101.15 and \$0 with equal probability over a certain payment of \$100.

[15] While we find this to be a reasonable assumption in this case, and approaches like this one are not rare in experimental economics, we recognize the exercise suffers from a "circularity trap," by which we identify bias by looking for choices that conflict with true preferences while inferring true preferences from unbiased choices (Bernheim and Taubinsky, 2018).

in part 1 and hiring the Good computer in part 2, and the other is a lottery resulting from choosing the Pooling task in part 1 and hiring a computer randomly in part 2. Figure A.1 in the appendix summarizes the results. Across all six parameterizations, less than 3% of participants choose the lottery induced by the Pooling task.

Parameterizations 7–10: The Pooling task produces a certain payoff that is strictly larger than the certain payoff produced by the Screening task, unambiguously constituting a mistake.

II.D. Baseline and Strategy Method Treatments

We use two treatments to show that participants do not screen optimally: the *Strategy Method* treatment, which establishes a benchmark for the amount of noise in the participants' answers, and the *Baseline* treatment, which identifies suboptimal screening beyond that which comes from our noise benchmark.[16]

***Baseline* treatment:** We design this treatment to mimic the natural order of choices in a naturalistic hiring scenario: The participants first choose the trial task, then move on to the hiring stage, where they observe the payoff each computer generates and choose which one to hire. As in any experiment, we expect some participants to make the mistake of choosing the Pooling task for irrelevant reasons, such as lack of attention and trembling hand errors, which we interpret as experimental noise that artificially increases the rate of mistakes. Thus, the mistake rate we observe in the *Baseline* treatment combines the screening mistakes caused by stable cognitive errors—which we are interested in—with mistakes stemming from noise. We separate the two to measure the true prevalence of screening mistakes using the *Strategy Method* treatment.

***Strategy Method* treatment:** We design this treatment to measure the share of mistakes that come from experimental noise. This control treatment is meant to help partici-

[16] For the *Strategy Method* to be a good benchmark, we assume that the noise in responses to the *Baseline* and *Strategy Method* treatments is similar.

pants as much as possible while keeping the structure of the experiment the same. We use the share of participants who still make suboptimal choices in such a scenario as a measure of noise. In this treatment, participants submit their whole strategy, which requires solving the problem backward. They start with the part 2 hiring decision and choose which computer to hire in each contingency, when the part 1 trial task is the Screening task (i.e., when they know the computers' quality) and when the part 1 trial task is the Pooling task (i.e., when they do not know the computers' quality).[17] Next, they complete the part 1 trial task. At this point, we remind them about their contingent decisions for the hiring stage. The participants also see the payoff consequences of their choices next to the two task options, which provides no new information but helps them conveniently assess the two options. Because the *Strategy Method* is such a strong intervention, we attribute any mistakes in this treatment to inherent noisiness and use it to establish the benchmark amount of noise in the experiment.[18]

II.E. Mechanisms: Automatic Inference and Plan Treatments

By our experimental design, the mistake of choosing the Pooling task cannot be attributed to standard explanations of time inconsistency, risk aversion or mistakes in strategic reasoning. To solve the problem correctly, participants must *recognize* and *use* the informational content in the computers' performance on the part 1 Screening task. We separately test for both — whether they fail to anticipatedly recognize the inference they will be able to make (henceforth, "Failure to Anticipate Inference"), and whether they fail to use it in planning their strategy (henceforth, "Failure to Plan"). To test Failure to Anticipate Inference, we design the *Automatic Inference* treatment, in which we tell the participants that we will make the inference for them. To test Failure to Plan, we design the *Plan* treatment, in which

[17] Specifically, in choosing for the contingency in which the part 1 trial task is the Screening task, the participants see the labels for the Good and Bad computers before making their hiring decision.

[18] Note that to the extent that this treatment is not strong enough to remove the cognitive bias that drives mistakes, the *Strategy Method* will overestimate noise. We can then interpret our estimates as lower bounds for the prevalence of the screening mistakes we study.

we make participants choose their full strategy from the beginning. These interventions substantially reduce the mistakes in Screening-optimal rounds, which supports Failure to Anticipate Inference and Failure to Plan as relevant mechanisms driving suboptimal underscreening.

The Failure to Anticipate Inference suggests that participants do not know that they will be able to make inferences in the future: they do not realize that the Screening task provides information that reveals each computer's quality. Despite this, they may be thinking about part 2 and planning their strategy for the future. The *Automatic Inference* treatment, a light-touch intervention to the *Baseline* treatment, tests for this mechanism.

***Automatic Inference* treatment:** This treatment is different from the *Baseline* treatment in one crucial way: instead of participants making the inference themselves, we make the inference automatically for them. If the participant chooses the Screening task, we explicitly tell them the quality of each computer before their hiring decision. If the participant chooses the Pooling task, we do not reveal the computer quality. Participants know this when choosing between the Screening and Pooling tasks in part 1. Specifically, next to the Screening task option in the experiment interface, we explain that we will reveal the computers' quality; next to the Pooling task option, we explain that we will not. If the hypothesized Failure to Anticipate Inference mechanism is correct, this additional message at the first stage should significantly reduce the mistake rate.

By our second hypothesized mechanism, Failure to Plan, participants do not plan their entire strategy when choosing the trial task; instead, they myopically focus on the immediate choice in front of them. We design the *Plan* treatment to test for Failure to Plan.

***Plan* treatment:** This treatment changes how we elicit participants' strategy relative to the *Baseline* treatment. We ask participants to choose a complete plan for their strategy from the beginning. Specifically, they choose between three options: (i) Choosing the Screening task and hiring the computer that produces the larger amount, (ii) choosing the Screening task and hiring the computer that produces the smaller amount, and (iii) choosing the Pooling task and hiring one the computers chosen randomly. By simultaneously choosing for parts 1 and 2, this elicitation forces participants to consider the whole problem. If the Failure to Plan mechanism is causing mistakes, this planning tool should reduce the rate of mistakes relative to the *Baseline* treatment.

II.F. Procedures: Online Experiments on Prolific

We recruited all participants on Prolific, an online platform frequently used for research studies, on October 10^{th} 2023 and April 25^{th} 2024. We restrict the sample to participants in the USA who are fluent in English and have completed at least 100 previous submissions on Prolific, with a minimum approval rate of 97%. The experiment was implemented using the oTree platform (Chen et al., 2016). The study was registered on the AEA RCT registry with ID AEARCTR-0012230 under the title "The Inference Cost of Interventions."

We recruited 982 participants who were randomly assigned to the four treatments: 251 subjects were assigned to the *Baseline* treatment, 244 to *Strategy Method*, 244 to *Automatic Inference*, and 251 to *Plan*.[19] Participants receive a $3 completion payment and an average bonus payment of $3.79, and the median completion time is 21 minutes, which is equivalent to $19.40 per hour.

We follow the standard procedures to ensure that the results are not driven by misunderstanding of the experiment instructions. Participants receive detailed instructions and must correctly answer a set of understanding questions about them before proceeding, as well as passing attention checks. Figure A.2 in the appendix shows the distribution of mistakes on the understanding questions.[20] Restricting the sample to participants who make zero mistakes does not affect the results.[21] Importantly, we present participants with parts

[19] The sample is balanced on gender. The average age is 42 years. 74% of the sample identify as White and 11% as Black. 53% of the sample have a Bachelor's degree or higher. Table A.1 in the appendix summarizes demographic characteristics by treatment.

[20] For example, participants cannot proceed in the experiment without correctly answering that the computers in part 1 and part 2 are the same.

[21] Additionally, after the *Baseline* participants make their part 1 decision, we ask them whether they will see how much money each computer produced in part 1. Around 71% of the participants answered this question

1 and 2 on the same page, so that it is unlikely that failures to consider part 2 into their part 1 decision is an artifact of the experiment (see appendix section I.B for screenshots of the experimental instructions).

III. Results: Prevalence of Mistakes

The *Baseline* and *Strategy Method* treatments show that people fail to account for the inference consequences of their policy choices. Most participants—68%—suboptimally choose the Pooling task, and the mistakes do not disappear entirely with experience. The *Automatic Inference* and *Plan* treatments substantially reduce the mistake rate, suggesting that the Failure to Anticipate Inference and Failure to Plan mechanisms are important drivers of insufficient screening. Participants do not recognize they can extract valuable information in the future, and do not plan the entire strategy.[22]

III.A. Baseline and Strategy Method Treatments

Choosing the Pooling task is a mistake in six of the ten parameterizations. Since we are interested in documenting insufficient screening, we focus on these six parameterizations in the rest of the analysis.[23] We expect participants to learn over time from feedback and experience. Therefore, we present the main results separately by round.

Figure II shows the main results. In the first round, most participants screen insufficiently: 68% of participants choose the Pooling task in the *Baseline* treatment. In the *Strategy Method* treatment, only 18% choose the Pooling task. The difference is statistically significant (t-test yields a p-value < 0.001). Overall, in the six Screening-optimal rounds,

correctly on the first try, confirming that most participants understand this essential aspect of the experimental design (restricting to only these participants does not affect the results, as we discuss in section IV).

[22] While we design each treatment to tackle one mechanism, we cannot rule out the possibility that either treatment allows participants to correct both mechanisms (for example, because the interaction of both mechanisms made it too complex for them to realize, but the elimination of one allows for a full correction).

[23] We include Pooling-optimal rounds to prevent participants from mechanically choosing the Screening task without regard for the payoffs. Hence, they serve an auxiliary role since there is no relevant trade-off in these rounds.

participants make an average of 2.38 and 1.23 mistakes in the *Baseline* and the *Strategy Method* treatments, respectively (p-value < 0.001).

Learning: There is substantial learning from experience, but mistakes disappear only partially. In rounds 2–10, the rate of mistakes reduces to an average of 34% in *Baseline*. It is still significantly higher than the rate of mistakes in the *Strategy Method* treatment, which is, on average, 21% (t-test yields a p-value < 0.001).[24] Most of the learning happens after round 1, with no learning in the *Strategy Method* treatment and little learning across rounds 2–10 in the *Baseline* treatment. The difference in the mistake rate between the *Baseline* and the *Strategy Method* treatments is statistically significant in nine out of ten rounds. Appendix Table A.2 further shows that mistakes are robust to learning by showing that the presence of interspersed pooling-optimal rounds does not disrupt the learning process; Facing screening-optimal rounds in a row does not substantially change the learning pattern, with the rate of mistakes in *Baseline* remaining steady at around 40%. Moreover, Appendix Table A.3 confirms most learning happens between the first and second rounds, with minimal learning as participants face their third, fourth, fifth or sixth screening-optimal round.[25]

III.B. Mechanism: Automatic Inference and Plan Treatments

The mistake rate in the *Automatic Inference* treatment is significantly lower than in the *Baseline* treatment. Figure II compares the mistake rates in all treatments. In the first round, only 44% of the participants make the mistake of choosing the Pooling task. This mistake rate is about halfway between the *Baseline* (68%) and the *Strategy Method* (18%) treatments. In subsequent rounds, we observe complete learning to the same level as in the *Strategy Method* treatment. The difference in the mistake rate from the *Baseline* is

[24] That mistakes in *Baseline* do not fully disappear with experience is consistent with the experimental evidence in Chakraborty and Kendall, 2022 and that reviewed in Niederle and Vespa, 2023 for failures of contingent thinking.

[25] While the rate of mistakes goes down in the last round for *Baseline*, the same happens in the benchmark *Strategy Method* treatment.

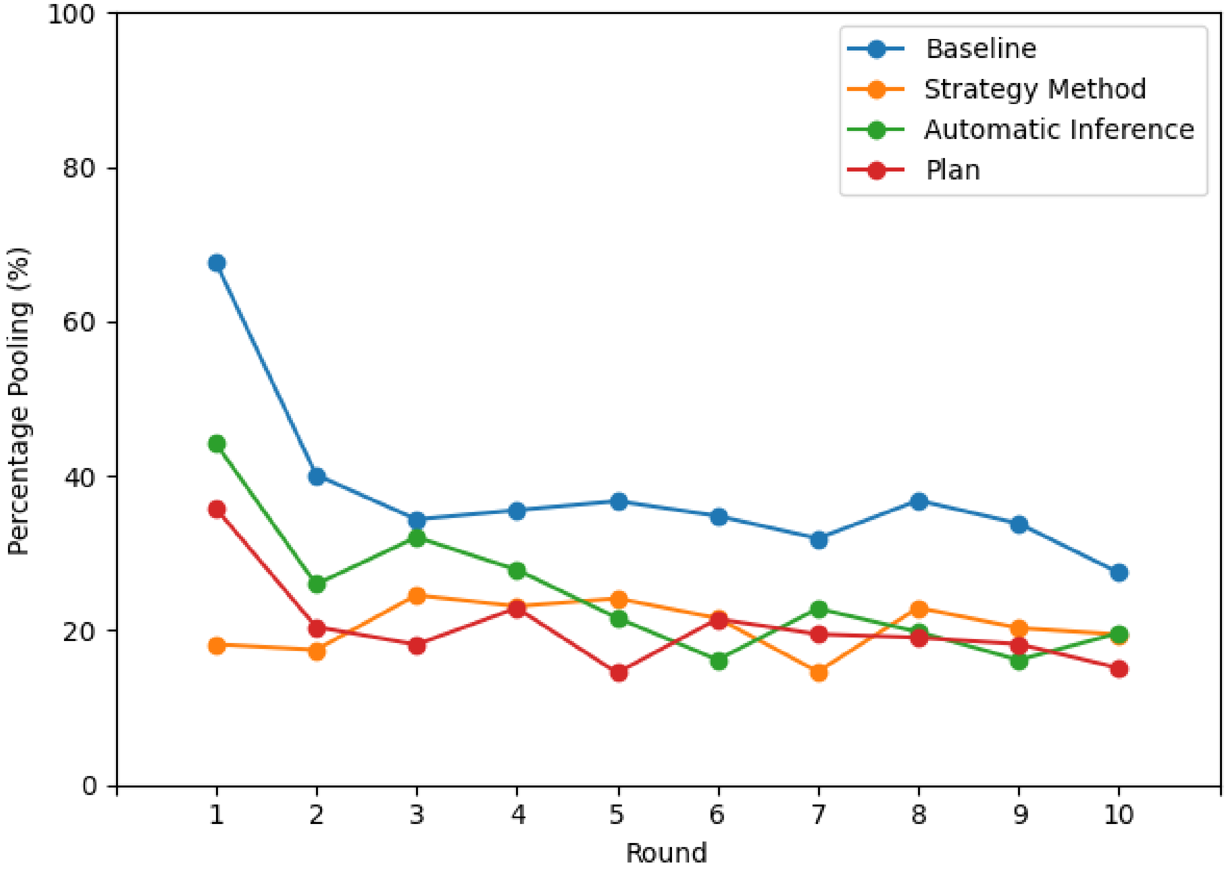


Figure II: Rate of mistakes across rounds with Screening-optimal parameterizations for all treatments.

statistically significant in round 1 and in rounds 2–10 combined (t-tests yield p-values < 0.001). Considering each round separately, the mistake rate is significantly different from the *Baseline* treatment in seven out of ten rounds. Thus, the *Automatic Inference* treatment suggests that the hypothesized Failure to Anticipate Inference mechanism is important: Participants seem not to understand how to infer information in the future.

Similarly, the mistake rate in the *Plan* treatment is significantly lower than in the *Baseline* treatment (t-tests yield p-values < 0.001). In the first round, 36% of the participants choose the Pooling task.[26] In subsequent rounds, participants learn completely — the mistake rate in rounds 2–10 is similar to the *Strategy Method* treatment.[27]

The difference in the mistake rate from the *Baseline* is significant in all rounds. Thus, the *Plan* treatment suggests that the hypothesized Failure to Plan mechanism is important too: Participants do not seem to plan their entire strategy.[28] This result is in line with evidence by Chakraborty and Kendall, 2022 supporting the idea that participants fail to think through one's own actions in the future, also in the context of an individual decision-problem.

All six Screening-optimal parameterizations considered, participants make an average of 1.57 and 1.29 mistakes in the *Automatic Inference* and *Plan* treatments, respectively, compared to 2.38 and 1.23 in the *Baseline* and the *Strategy Method* treatments.[29] The difference in the average number of mistakes between treatments hides heterogeneity in the distribution of mistakes. For each participant, we calculate the number of mistakes they make in the Screening-optimal rounds (i.e., choosing the Pooling task). Aggregating over all participants within a given treatment, we plot the histograms of the total number of mistakes in

[26] The difference between *Automatic Inference* and *Plan* treatments is not statistically significant at the 5% level.

[27] Appendix Table A.3 confirms this learning pattern when focusing on the number of screening-optimal round the participants have faced.

[28] Appendix Table A.2 suggests the *Automatic Inference* treatments becomes less effective at reducing mistakes as participants see more screening-optimal rounds in a row, yet this exercise is much less well-powered, making it hard to draw reliable inferences.

[29] Relative to the *Strategy Method*, the difference is statistically significant for the *Automatic Inference* treatment (p-value < 0.001), but not for the *Plan* treatment (p-value = 0.5).

Appendix Figure A.3. In contrast to the other three treatments, a disproportionately large group of participants in the *Baseline* treatment always make the mistake of choosing the Pooling task, and a disproportionately small group of participants never make this mistake.

Participants in the *Baseline* treatment who always choose the Pooling task drive the persistence of insufficient screening. These participants never learn that screening is optimal and consistently avoid screening throughout the experiment. If we remove these participants from the sample, round 1 results remain similar, but the mistake rate in rounds 2–10 becomes close to statistically indistinguishable in all treatments except *Plan* (see column 7 in Table II). This result shows that some participants struggle to learn the optimal screening choice without aid, making the mistake rate in the *Baseline* treatment more persistent.[30]

This section's results suggest that participants fail to anticipate inference and plan their strategies. These two mechanisms seem to be similarly important drivers of insufficient screening.

IV. Robustness of Results

Table II summarizes the multiple tests we conduct to verify the robustness of our results.

Pooling is a mistake. We directly elicit preferences over the induced lotteries, and almost all participants choose the lottery corresponding to the Screening task. Nevertheless, a few participants sometimes choose the lottery corresponding to the Pooling task, which could explain our results. To rule out this possibility, we construct a subsample without them and re-estimate the magnitude of mistakes. Column (2) of Table II shows that excluding participants who prefer at least one lottery induced by the Pooling task does not affect our results.

Participants' understanding. Throughout the experiment, participants had to correctly answer a set of understanding questions. If they made an error on an understanding

[30] Participants who always choose the Pooling task never observe the informational gains from screening. Footnote 33 discusses an alternative treatment that can test for the effect of forced exposure to the gains from screening.

question, they had to correct it before proceeding. This design helps ensure that participants understand the critical features of the experiment even if they missed them while reading instructions.

Appendix Figure A.2 plots the distribution of errors in understanding questions. Most participants make few errors, with 89% of them making at most one error, and 56% making zero errors. The small number of understanding errors highlights that confusion or misunderstanding was not an issue for most participants. Since error rates are similar across treatments, misunderstanding is unlikely to drive the treatment effects.[31] The results are robust to misunderstanding by focusing on the 546 participants who make zero errors on the understanding questions, as Column (3) of Table II shows.

Do participants in the *Baseline* treatment understand that they will observe how much each computer produces before they make the hiring choice? If they fail to understand this, the Screening task is not necessarily optimal. However, asking an understanding question about this at the beginning of the experiment would highlight the connection between the two parts of the problem, potentially shutting down the mechanisms that we are interested in studying. To avoid interfering with the mechanisms, we first elicit participants' round 1 trial task choice and only then ask the understanding question about whether they will observe how much each computer produces.[32] Restricting to the 71% of participants who answer this understanding question correctly—they know that they will observe how much each computer produces before the hiring choice—does not affect the results, as we show in column (4) of Table II.

Optimal use of information. Screening is optimal only if the participants correctly use the revealed information. To claim that choosing the Pooling task is a mistake, we need to show that participants who do so would be able to use the information if they got it. An

[31] The average number of errors in the *Baseline* treatment is significantly smaller than in the *Plan* treatment and not significantly different from the *Strategy Method* and *Automatic Inference* treatments. The distributions of errors are similar across treatments (Appendix Figure A.4).

[32] As a result, the understanding question does not affect round 1, but it may speed up learning, reducing the mistake rate in rounds 2–10 in *Baseline*, which goes against our hypothesis.

ideal test for this would be observing participants' counterfactual part 2 hiring choices, after the part 1 Screening task. We approximate this counterfactual scenario by looking at participants' actual part 2 hiring choices after choosing the Screening task in part 1. All Screening-optimal parameterizations are similar, with the only differences being parameters and task names. Therefore, we treat rounds where a participant chooses the Screening task as a counterfactual for the rounds where they choose the Pooling task. Across all treatments, more than 90% of participants hire the Good computer after choosing the Screening task (Appendix Figure A.5). This suggests that it is unlikely that participants who do not screen would not be able to use the revealed information.

Although mistakes at the hiring stage differ across treatments, this does not drive the differences in the screening decision. Participants in the *Baseline* treatment make slightly more hiring mistakes than in the other treatments. These differences are too small to explain the large differences in the trial task choices. Nevertheless, we verify that the results are robust to this concern by restricting the sample to those participants who use the information correctly: those who chose the Screening task at least once and always hired the Good computer thereafter. Round 1 results stay the same in this subsample as in the whole sample, although the levels of mistake rates mechanically decrease for all treatments. In subsequent rounds, most treatment effects disappear—participants in the *Baseline* treatment make the same number of mistakes as in the *Strategy Method* and *Automatic Inference* treatments, although still more than in the *Plan* treatment (Table II, column (6)). Both, the decrease in the levels of mistakes and the faster learning dynamics are expected because this subsample excludes everyone who always chooses the Pooling task, who drive the persistence of insufficient screening in the full *Baseline* sample.[33]

[33] An alternative approach to test for the optimal use of information among those who do not screen is to exogenously reverse the trial task choice from Pooling to Screening, to randomly force participants who choose Pooling to face the consequences of picking the Screening task. Early pilot results show that most participants hire the Good computer after the trial task choice is reversed from Pooling to Screening. We chose not to implement this approach in the main experiment for two reasons. First, pilot participants report high levels of confusion stemming from this random exogenous reversal, which could drive mistakes. Second, it provides a counterfactual only probabilistically, in a subset of rounds, which would demand a larger sample size.

Stakes: The payoff consequences of these mistakes are substantial. By not screening, participants leave on the table at least \$2, which represents 53% of their average bonus payment. While whether these mistakes disappear as stakes increase remains an empirical question, and we do not expect experienced policy-makers or employers to necessarily behave as our Prolific participants, the sample and stakes we study are important in themselves in shedding light on lay people's behavior in the face of this trade-off, which can determine their support for policies.

First round learning: Many naturally occurring decision problems are effectively one-shot or first-time events. Individuals often lack repeated exposure to such situations, limiting learning opportunities.[34] Consequently, behavior in the initial round of the experiment can be highly informative, as it mirrors the conditions under which many naturally occurring screening decisions are made. Even if participants later learned, their first-round behavior provides valuable insights into how individuals approach unfamiliar trade-offs in the absence of prior experience.

Nevertheless, it is crucial to understand the persistence of the mistake, and the substantial and immediate drop in the rate of mistakes after the first round calls for caution in interpreting their persistence in our setting. As we argue in our discussion of participants' understanding, the rate of mistakes does not seem to be driven by confusion or lack of engagement. However, the fact that behavior changes so substantially after round 1, sheds light on the persistence and robustness of the cognitive bias.

We might be estimating an upper bound in learning: As we discuss above, after round 1 we introduce a comprehension question meant to test whether participants in the *Baseline* treatment understand that they will observe how much each computer produces before they make the hiring choice. This question highlights the connection between the two parts of the problem, potentially shutting down the mechanisms that we are interested in study-

[34] For example, a hiring manager deciding whether to invest additional resources in screening job applicants may face only a few such high-stakes hiring decisions in their career. Similarly, individuals choosing whether to obtain a second medical opinion before a major procedure often do so without prior experience in comparable situations.

ing, which is why we only ask it after round 1. As a result, the understanding question does not affect round 1, but it may speed up learning, leading to an underestimation of the persistence of the mistake.

Moreover, while mistakes exhibit significant persistence, there is heterogeneity in the learning pattern. In contrast to the substantial share of participants who make a mistake only in the first round, Figure A.3 shows that a disproportionately large group of participants always make the mistake in *Baseline* relative to the other treatments. This heterogeneity suggests that the robustness of the cognitive failure is not globally robust but might rather vary across individuals.

Individual characteristics. We rule out that differences in participants' self-reported age, gender, race, and education across samples drive the treatment effects. The treatments are balanced on these demographics but participants in the *Plan* treatment are slightly younger than the *Baseline* participants (Appendix Table A.1). Column (5) of Table II shows the results are robust to controlling for these characteristics.[35] Moreover, we find no evidence that the heterogeneity in the persistence of the mistake that we note in our discussion of learning after round 1 is related to self-reported age, gender, race, or education.

V. Conclusion

We experimentally show that failures of contingent thinking lead people to not screen enough when screening is optimal, even without time inconsistency, risk aversion, and strategic considerations. The rate of insufficient screening exceeds the rate at which participants make other mistakes when facing screening trade-offs, like screening when pooling is optimal. Suboptimal screening significantly diminishes, yet persists, with experience and feedback. We further postulate and test two mechanisms that drive insufficient screening. First, participants fail to recognize that screening allows them to make inferences in the future. Sec-

[35] Controls for age, race and "other gender," not shown in column (5) of Table II, are significantly different from zero (e.g., older and black participants are more likely to make mistakes).

Table II: Summary of estimates

Panel A: Round 1 estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Strategy Method	-0.50***	-0.51***	-0.53***	-0.49***	-0.49***	-0.49***	-0.47***
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)
Automatic Inference	-0.23***	-0.25***	-0.27***	-0.23***	-0.23***	-0.27***	-0.21***
	(0.04)	(0.04)	(0.06)	(0.05)	(0.04)	(0.05)	(0.05)
Plan	-0.32***	-0.34***	-0.39***	-0.31***	-0.32***	-0.30***	-0.30***
	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)
Baseline mistake rate	0.68	0.68	0.66	0.67	0.68	0.63	0.62
N	982	934	546	909	982	742	914

Panel B: Rounds 2–10 estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Strategy Method	-0.13***	-0.14***	-0.15***	-0.14***	-0.13***	-0.02	-0.05**
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Automatic Inference	-0.11***	-0.12***	-0.11***	-0.12***	-0.11***	-0.04	-0.05*
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Plan	-0.15***	-0.17***	-0.17***	-0.16***	-0.15***	-0.07***	-0.09***
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.02)
Baseline mistake rate	0.34	0.34	0.29	0.35	0.34	0.20	0.23
N	4910	4670	2730	4545	4910	3710	4570

OLS regression results with binary dummies for each treatment (*Baseline* omitted) and mistakes in screening-optimal rounds as dependent variable. Specification by column: (1): Main sample, (2): Participants who never prefer the lottery induced by the Pooling task, (3): Participants who make zero mistakes on the understanding questions, (4): Participants who know that they will observe how much computers produce, (5): Main sample with controls for age, gender, race, and education level, (6): Participants who make zero mistakes in part 2 hiring choice, (7): Participants who choose the Screening task when it is optimal at least once. Robust standard errors in parentheses for round 1 and clustered by participant for rounds 2–10. $p-value < 0.1$: *, $p-value < 0.05$: **, $p-value < 0.01$: ***

ond, they fail to plan their entire strategy. We run two additional treatments that shut down each of the mechanisms, each of which accounts for about half of the non-noise mistake rate.

Our results suggest that individuals may fail to consider screening effects when choosing incentives and designing or supporting policies. Policies that create strong incentives induce uniform behavior, which limits the amount of information that can be learned from observed behavior. Thus, a trade-off exists between creating strong incentives and losing valuable information. The inference cost of policy choices is a crucial yet usually overlooked factor.

Consider an illustrative example: if prisons are strict about enforcing good behavior, every inmate will follow the rules. This policy may reduce violence within prisons, but it also prevents us from learning about inmates' qualities and their willingness to follow the law. Learning such information can be essential for early parole decisions, but it is lost if policies are so strict that everyone behaves the same.

While we cannot irrefutably argue that any particular policy choice is a mistake because of its inference cost—such a claim would require careful empirical analysis beyond the scope of this paper—, to the extent that we identify people screen insufficiently, our results suggest that optimal policies should allow, on the margin, for more screening opportunities.

We propose several avenues for future research. First, in our experiment, participants make inferences about computers. While this allows us to keep control, an exciting extension would be considering inferences about other people or *the self* in more realistic scenarios. For example, inferences about *the self* play a role in the literature on the unintended consequences of policies (Bitler and Karoly, 2015; Nandi and Laxminarayan, 2016), which shows that external incentives may crowd out intrinsic or 'warm glow' motivation (Lepper et al., 1973; Frey and Jegen, 2000; Bénabou and Tirole, 2006; Ariely et al., 2009). Our findings imply that policy supporters may not realize that strong incentives limit people's ability to feel good about themselves based on inferences they can make about themselves, leading to the crowding-out effect. A second example is the literature that shows the persistence of stereotypes and wrong beliefs (Snyder, 1981; Skrypnek and Snyder, 1982; Babcock et al., 2017). One mechanism for this persistence can be the existence of strong incentives for uniform behavior by which the stereotyped group cannot disprove the stereotype, leading to its persistence. Future work testing for the Failure to Anticipate Inference and to Plan in these settings would help understand the psychological underpinnings of these results.

As with any economic experiment, the introduction of a highly controlled environment comes at the cost of naturalism. While we believe our design captures the essence of the screening decision in an externally relevant way, future field work testing the role of contingent thinking in screening can further contribute to our understanding of this phenomenon. For example, while in naturally occurring scenarios screening is often intentional (e.g., hiring), the inference cost of policies can often be overlooked in situations in which preventing bad behavior takes a more salient role (e.g., prison policy). As another example, the external relevance of our results could be further tested by empirically quantifying the inference costs of policies in naturally occurring settings where the party making the inference differs from the one making the screening choice (e.g., as in the example above with prisons and inmates), including contexts where the policy choice is paternalistic. Future work could explore how this and other naturally occurring factors—that our experiment controls for—interact with our results.

Finally, the basic cognitive bias we explore can, in principle, relate to any setting in which there is a risky choice that provides information. More work in this area is needed to figure out whether the cognitive phenomenon we study in this paper extends to other settings, like canonical search problems, stopping problems, and beyond. As an example of the breadth of settings that are related, Bernheim and Whinston, 1998 show that incomplete contracts may be optimal if they allow learning more about the counterparty than complete contracts. Future work could explore whether the Failure to Anticipate Inference and to Plan can lead people to ignore this channel and write too restrictive contracts.

References

ANDERSON, C. M. (2012): "Ambiguity aversion in multi-armed bandit problems," *Theory and decision*, 72, 15–33.

ARIELY, D., A. BRACHA, AND S. MEIER (2009): "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially," *American economic review*, 99, 544–55.

BABCOCK, L., M. P. RECALDE, L. VESTERLUND, AND L. WEINGART (2017): "Gender differences in accepting and receiving requests for tasks with low promotability," *American Economic Review*, 107, 714–747.

BANOVETZ, J. M. (2020): "Three Essays on Experimental Economics and Applied Microeconomics," .

BÉNABOU, R. AND J. TIROLE (2006): "Incentives and prosocial behavior," *American economic review*, 96, 1652–1678.

BERGEMANN, D. AND J. VÄLIMÄKI (2018): "Bandit problems," in *The new Palgrave dictionary of economics*, Springer, 665–670.

BERNHEIM, B. D. AND D. TAUBINSKY (2018): "Behavioral public economics," *Handbook of behavioral economics: Applications and Foundations 1*, 1, 381–516.

BERNHEIM, B. D. AND M. D. WHINSTON (1998): "Incomplete contracts and strategic ambiguity," *American Economic Review*, 902–932.

BHARGAVA, S., G. LOEWENSTEIN, AND J. SYDNOR (2017): "Choose to lose: Health plan choices from a menu with dominated option," *The Quarterly Journal of Economics*, 132, 1319–1372.

BINMORE, K., J. MCCARTHY, G. PONTI, L. SAMUELSON, AND A. SHAKED (2002): "A backward induction experiment," *Journal of Economic theory*, 104, 48–88.

BITLER, M. P. AND L. A. KAROLY (2015): "Intended And Unintended Effects Of The War On Poverty: What Research Tells Us And Implications For Policy," *J. Policy Anal. Manage.*, 34, 639–696.

CALFORD, E. M. AND T. N. CASON (2024): "Contingent reasoning and dynamic public goods provision," *American Economic Journal: Microeconomics*, 16, 236–266.

CHAKRABORTY, A. AND C. W. KENDALL (2022): "Noisy foresight," Tech. rep., National Bureau of Economic Research.

CHEN, D. L., M. SCHONGER, AND C. WICKENS (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

DAL BÓ, E., P. DAL BÓ, AND E. EYSTER (2018): "The demand for bad policy when voters underappreciate equilibrium effects," *The Review of Economic Studies*, 85, 964–998.

DUFWENBERG, M. AND M. VAN ESSEN (2018): "King of the hill: Giving backward induction its best shot," *Games and Economic Behavior*, 112, 125–138.

ELLSBERG, D. (1961): "Risk, ambiguity, and the Savage axioms," *The quarterly journal of economics*, 75, 643–669.

ESPONDA, I. AND D. POUZO (2017): "Conditional retrospective voting in large elections," *American Economic Journal: Microeconomics*, 9, 54–75.

ESPONDA, I. AND E. VESPA (2014): "Hypothetical thinking and information extraction in the laboratory," *American Economic Journal: Microeconomics*, 6, 180–202.

——— (2024): "Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory," *Review of Economic Studies*, 91, 2806–2831.

EYSTER, E. (2019): "Errors in strategic reasoning," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2, 187–259.

EYSTER, E. AND M. RABIN (2005): "Cursed equilibrium," *Econometrica*, 73, 1623–1672.

FREY, B. S. AND R. JEGEN (2000): "Motivation Crowding Theory: A Survey of Empirical Evidence, REVISED VERSION," *Working paper series/Institute for Empirical Research in Economics*.

HOELZEMANN, J. AND N. KLEIN (2021): "Bandits in the Lab," *Quantitative Economics*, 12, 1021–1051.

HUDJA, S. AND D. WOODS (2024): "Exploration versus exploitation: A laboratory test of the single-agent exponential bandit model," *Economic Inquiry*, 62, 267–286.

JOHNSON, E. J., C. CAMERER, S. SEN, AND T. RYMON (2002): "Detecting failures of backward induction: Monitoring information search in sequential bargaining," *Journal of economic theory*, 104, 16–47.

KWON, O. (2020): "Strategic Experimentation with Uniform Bandit: An Experimental Study," *Unpublished Manuscript*.

LEPPER, M. R., D. GREENE, AND R. E. NISBETT (1973): "Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis." *Journal of Personality and social Psychology*, 28, 129.

LEVITT, S. D., J. A. LIST, AND S. E. SADOFF (2011): "Checkmate: Exploring backward induction among chess players," *American Economic Review*, 101, 975–990.

MARTÍNEZ-MARQUINA, A., M. NIEDERLE, AND E. VESPA (2019): "Failures in contingent reasoning: The role of uncertainty," *American Economic Review*, 109, 3437–74.

MERLO, A. AND A. SCHOTTER (1999): "A surprise-quiz view of learning in economic experiments," *Games and Economic Behavior*, 28, 25–54.

——— (2003): "Learning by not doing: an experimental investigation of observational learning," *Games and Economic Behavior*, 42, 116–136.

MILGROM, P. AND J. ROBERTS (1986): "Relying on the information of interested parties," *The RAND Journal of Economics*, 18–32.

NANDI, A. AND R. LAXMINARAYAN (2016): "The unintended effects of cash transfers on fertility: evidence from the Safe Motherhood Scheme in India," *J. Popul. Econ.*, 29, 457–491.

NIEDERLE, M. AND E. VESPA (2023): "Cognitive limitations: Failures of contingent thinking," *Annual Review of Economics*, 15, 307–328.

SKRYPNEK, B. J. AND M. SNYDER (1982): "On the self-perpetuating nature of stereotypes about women and men," *Journal of Experimental Social Psychology*, 18, 277–291.

SLIVKINS, A. ET AL. (2019): "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, 12, 1–286.

SNYDER, M. (1981): "On the self-perpetuating nature of social stereotypes," *Cognitive processes in stereotyping and intergroup behavior*, 183.

TERGIMAN, C. J. (2024): "Correlation Neglect in Student-to-School Matching," *American Economic Journal: Microeconomics, forthcoming*.

A. Appendix

A.A. Tables and Figures

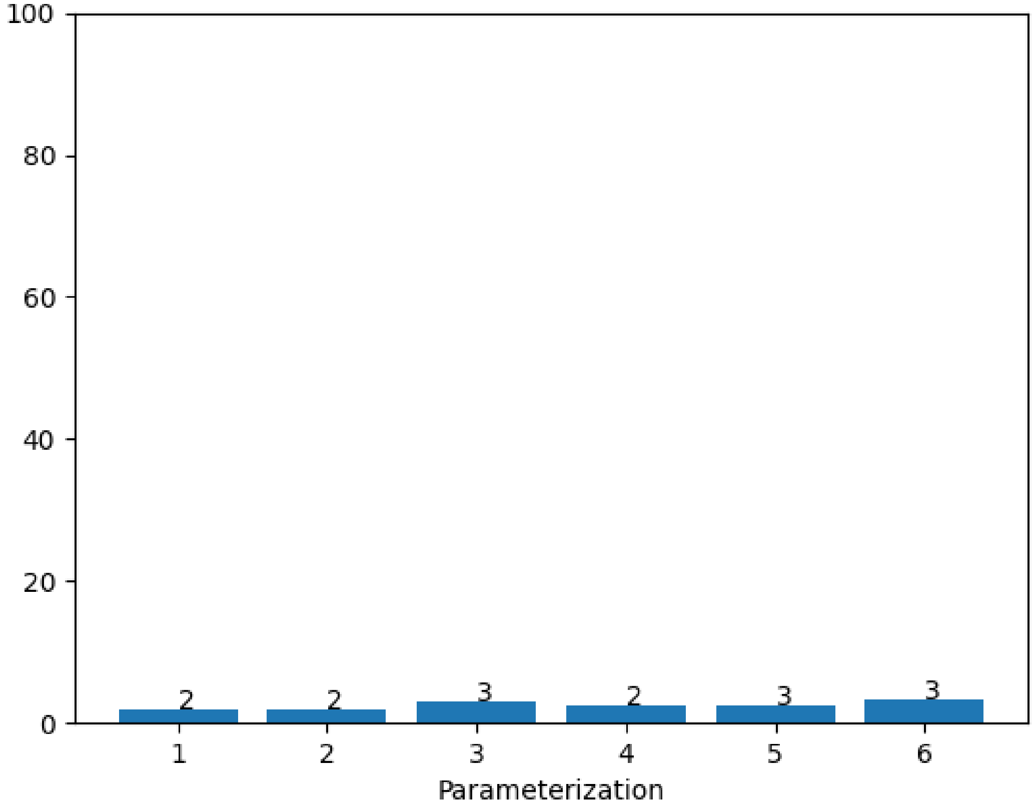


Figure A.1: The share of participants who choose the lottery induced by the Pooling task

Treatment	Age	p-val	Female, %	p-val	White, %	p-val	College, %	p-val
Baseline	43.17		52.99		73.31		50.60	
Strategy Method	43.72	0.65	49.15	0.40	77.54	0.28	58.05	0.10
Automatic Inference	41.87	0.27	47.13	0.19	74.18	0.83	56.56	0.18
Plan	40.84	0.04	49.00	0.37	71.31	0.62	48.21	0.59

Table A.1: Demographic characteristics (average age, percent female, and percent White) and share of college-educated by treatment. p-values are from t-tests for the difference from the *Baseline* treatment.

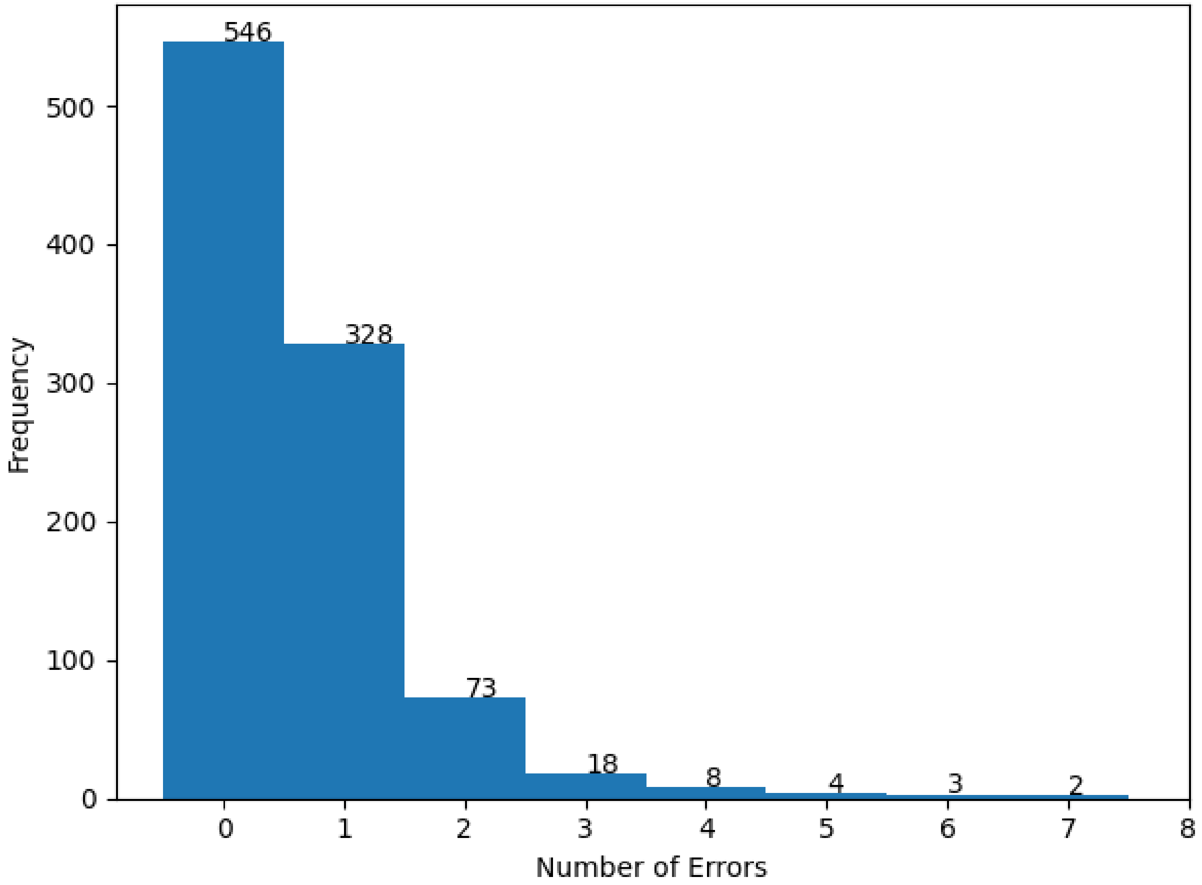


Figure A.2: The distribution of the number of errors on the understanding questions, pooling all treatments. Total sample size is 982.

	(1)	(2)	(3)	(4)	(5)
Strategy Method	-0.5***	-0.23***	-0.11	-0.23	-0.4*
	(0.04)	(0.06)	(0.09)	(0.14)	(0.22)
Automatic Inference	-0.23***	-0.14**	-0.02	-0.07	-0.07
	(0.04)	(0.06)	(0.09)	(0.15)	(0.29)
Plan	-0.32***	-0.2***	-0.25***	-0.37***	-0.4*
	(0.04)	(0.06)	(0.08)	(0.11)	(0.22)
Baseline mistake rate	0.68	0.4	0.35	0.37	0.4
N	982	491	215	63	17

Table A.2: Screening-optimal rounds in a row. OLS regression results with binary dummies for each treatment (*Baseline* omitted) and mistakes in screening-optimal rounds as dependent variable. Specification by column: (1): Round 1 for all observations, (2)-(5) Round [x] for observations where [x] first rounds were separating-optimal ([x] ranges from 2 to 5). For example, column (5) shows behavior in the fifth (screening-optimal) round for participants who faced four screening-optimal rounds in a row immediately before.

Table A.3: Estimates by Screening optimal round count

	(1)	(2)	(3)	(4)	(5)	(6)
Strategy Method	-0.5***	-0.15***	-0.13***	-0.16***	-0.13***	-0.08**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Automatic Inference	-0.23***	-0.12***	-0.1**	-0.13***	-0.14***	-0.08**
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Plan	-0.32***	-0.16***	-0.15***	-0.16***	-0.18***	-0.12***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Baseline mistake rate	0.68	0.37	0.36	0.35	0.34	0.27
N	982	982	982	982	982	982

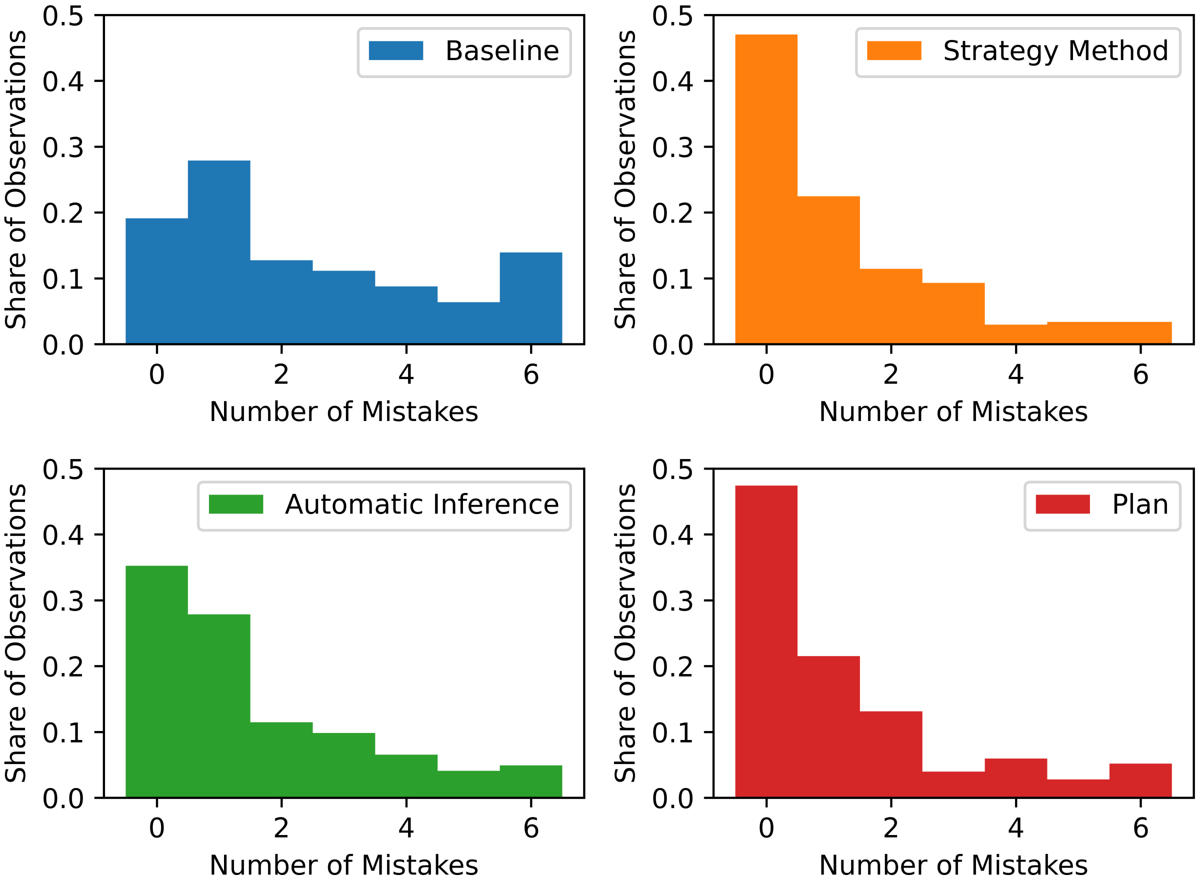


Figure A.3: The distribution of total screening mistakes, by treatment

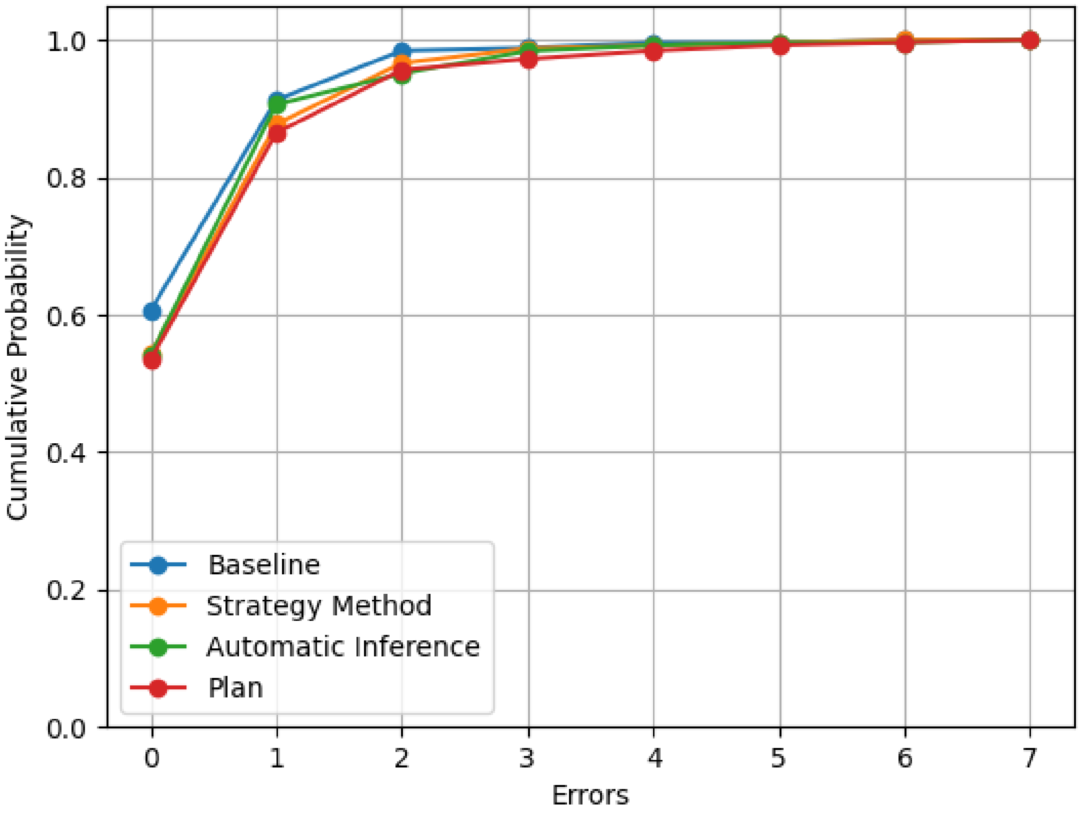


Figure A.4: The CDF of total errors on the understanding questions participants make by treatment.

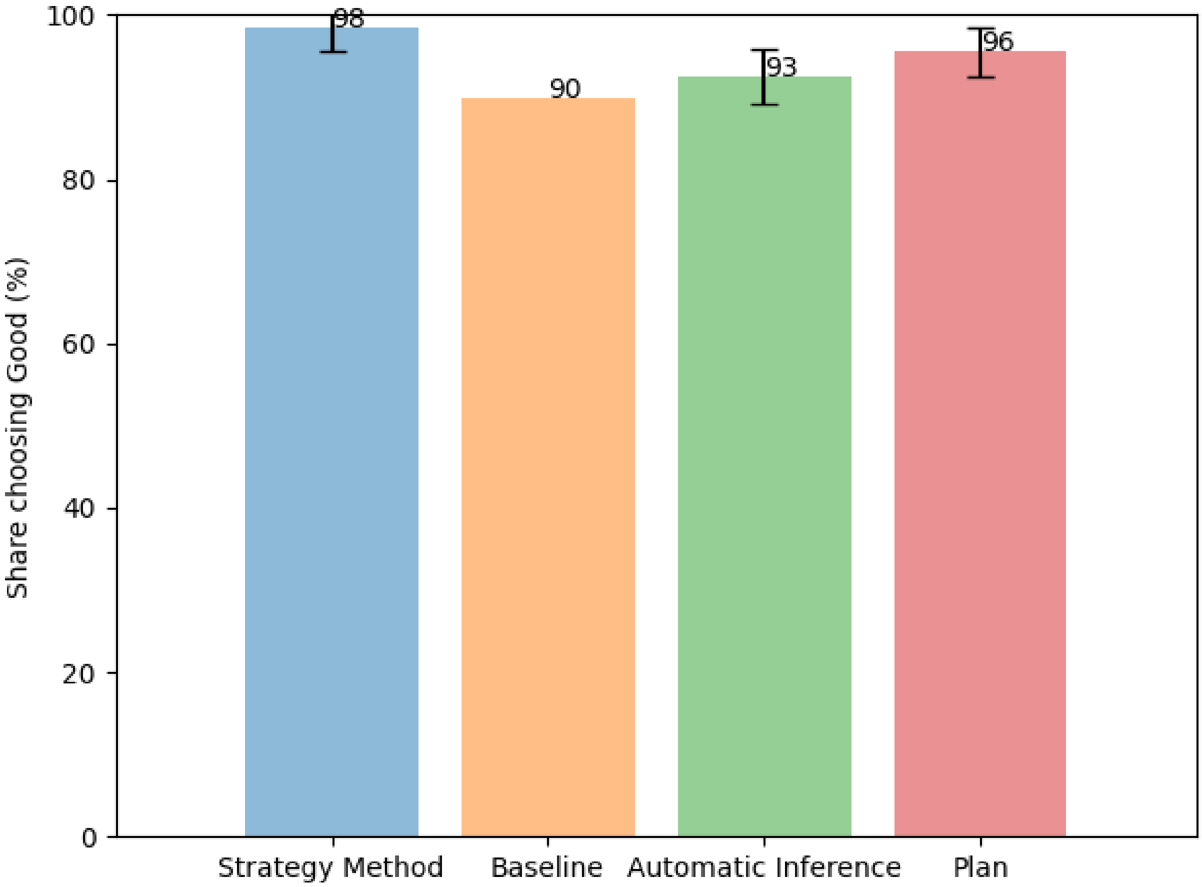


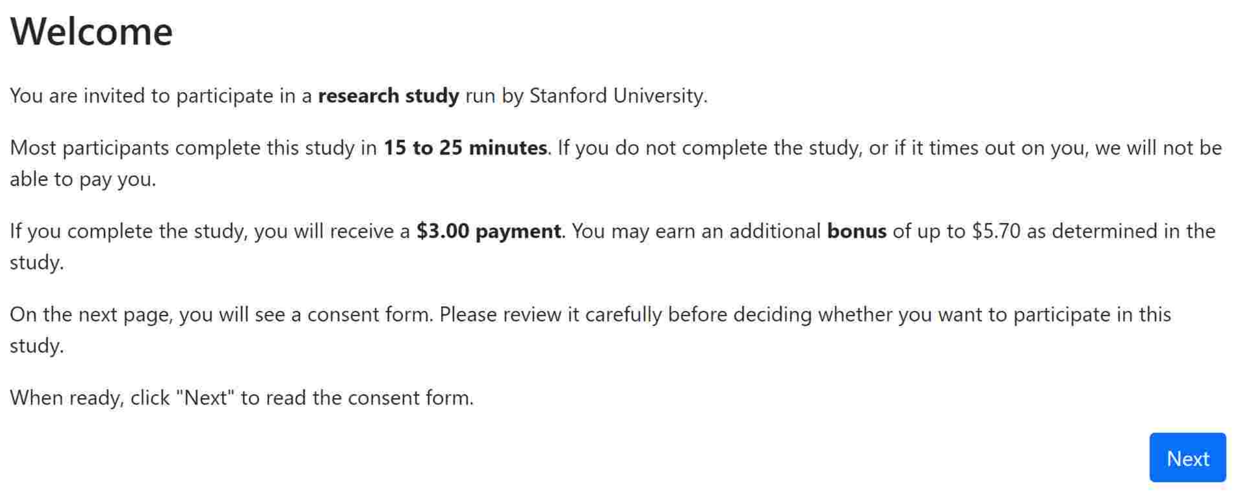
Figure A.5: The rate of hiring the Good computer in part 2 after choosing the Screening task by treatment.

A.B. Screenshots

This subsection presents the screenshots of the experiment for all four treatments in the order of their appearance. Pages that are identical for all treatments are grouped together. The screenshots are structured in the following way:

1. Introduction
2. Main decision—only round 1 (other rounds differ only in numbers and labels)
   (a) *Baseline* treatment
   (b) *Strategy Method* treatment
   (c) *Automatic Inference* treatment
   (d) *Plan* treatment
3. Other elicitations

Introduction



Welcome

You are invited to participate in a **research study** run by Stanford University.

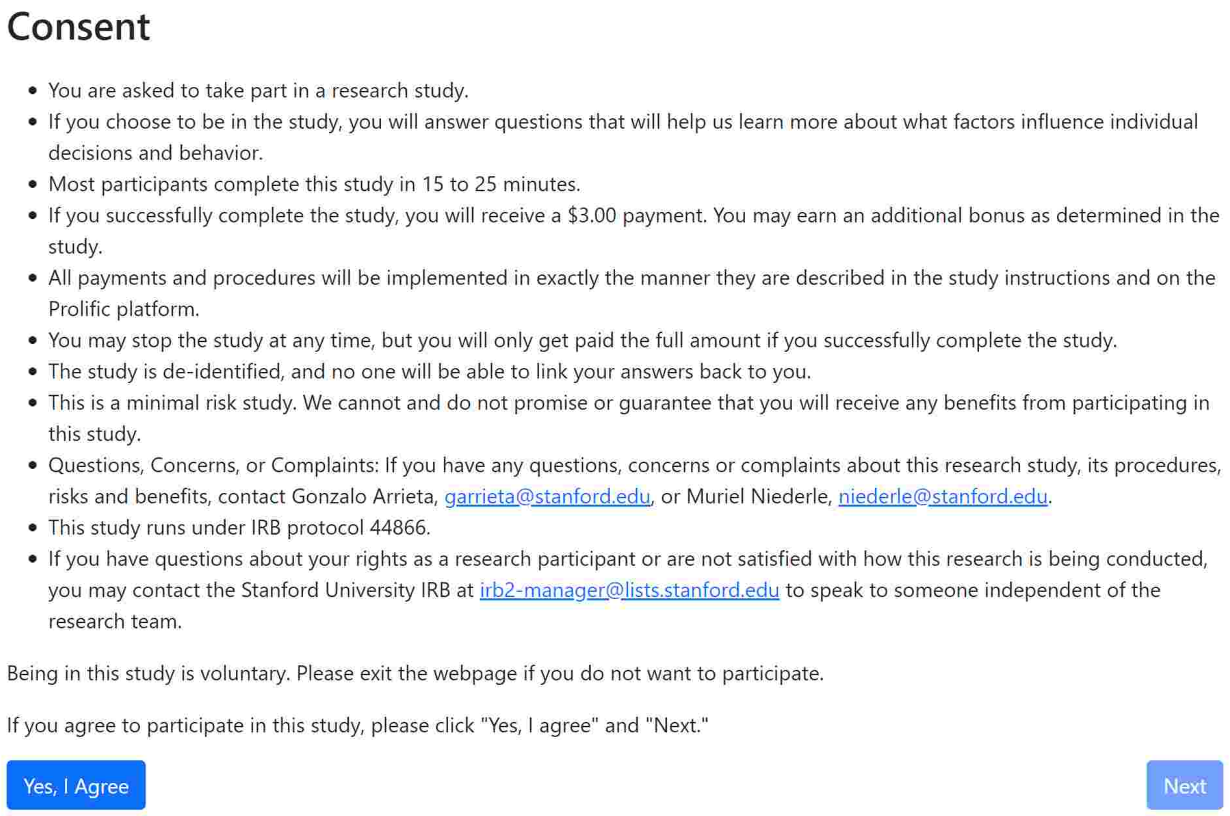
Most participants complete this study in **15 to 25 minutes**. If you do not complete the study, or if it times out on you, we will not be able to pay you.

If you complete the study, you will receive a **$3.00 payment**. You may earn an additional **bonus** of up to $5.70 as determined in the study.

On the next page, you will see a consent form. Please review it carefully before deciding whether you want to participate in this study.

When ready, click "Next" to read the consent form.

Figure A.6
Note: All treatments.



Consent

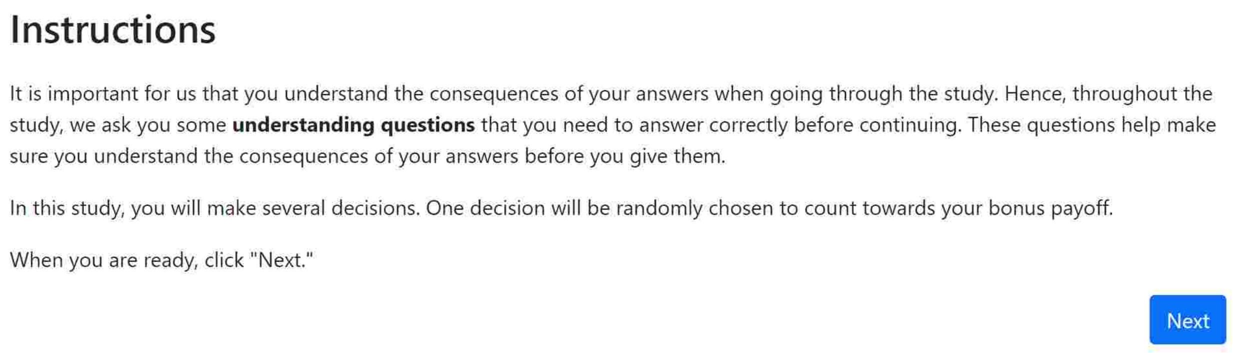
- You are asked to take part in a research study.
- If you choose to be in the study, you will answer questions that will help us learn more about what factors influence individual decisions and behavior.
- Most participants complete this study in 15 to 25 minutes.
- If you successfully complete the study, you will receive a $3.00 payment. You may earn an additional bonus as determined in the study.
- All payments and procedures will be implemented in exactly the manner they are described in the study instructions and on the Prolific platform.
- You may stop the study at any time, but you will only get paid the full amount if you successfully complete the study.
- The study is de-identified, and no one will be able to link your answers back to you.
- This is a minimal risk study. We cannot and do not promise or guarantee that you will receive any benefits from participating in this study.
- Questions, Concerns, or Complaints: If you have any questions, concerns or complaints about this research study, its procedures, risks and benefits, contact Gonzalo Arrieta, garrieta@stanford.edu, or Muriel Niederle, niederle@stanford.edu.
- This study runs under IRB protocol 44866.
- If you have questions about your rights as a research participant or are not satisfied with how this research is being conducted, you may contact the Stanford University IRB at irb2-manager@lists.stanford.edu to speak to someone independent of the research team.

Being in this study is voluntary. Please exit the webpage if you do not want to participate.

If you agree to participate in this study, please click "Yes, I agree" and "Next."

Figure A.7

Note: All treatments.



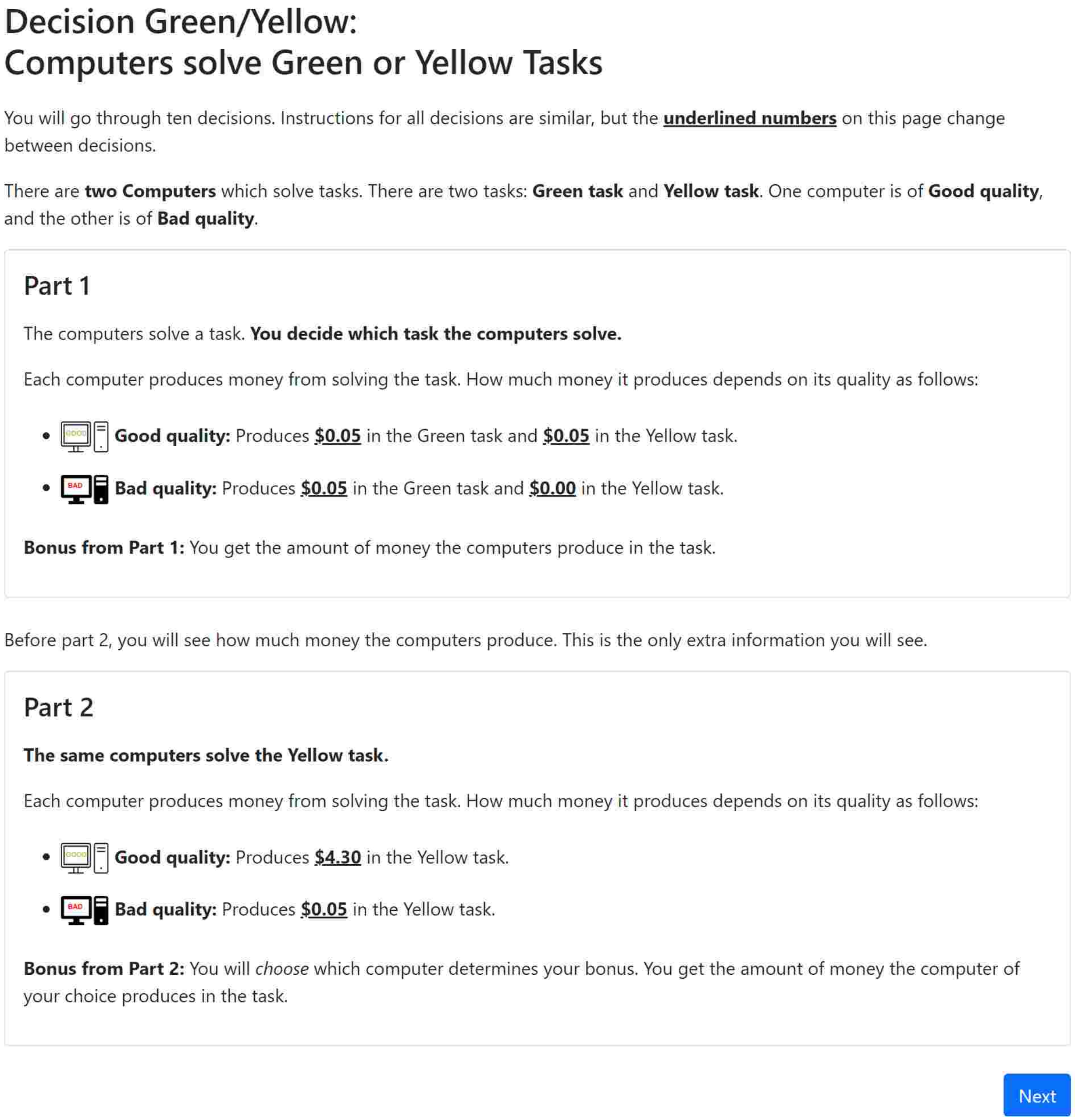
Instructions

It is important for us that you understand the consequences of your answers when going through the study. Hence, throughout the study, we ask you some **understanding questions** that you need to answer correctly before continuing. These questions help make sure you understand the consequences of your answers before you give them.

In this study, you will make several decisions. One decision will be randomly chosen to count towards your bonus payoff.

When you are ready, click "Next."

Figure A.8
Note: All treatments.



Decision Green/Yellow:
Computers solve Green or Yellow Tasks

You will go through ten decisions. Instructions for all decisions are similar, but the **underlined numbers** on this page change between decisions.

There are **two Computers** which solve tasks. There are two tasks: **Green task** and **Yellow task**. One computer is of **Good quality**, and the other is of **Bad quality**.

Part 1

The computers solve a task. **You decide which task the computers solve.**

Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$0.05** in the Green task and **$0.05** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Green task and **$0.00** in the Yellow task.

Bonus from Part 1: You get the amount of money the computers produce in the task.

Before part 2, you will see how much money the computers produce. This is the only extra information you will see.

Part 2

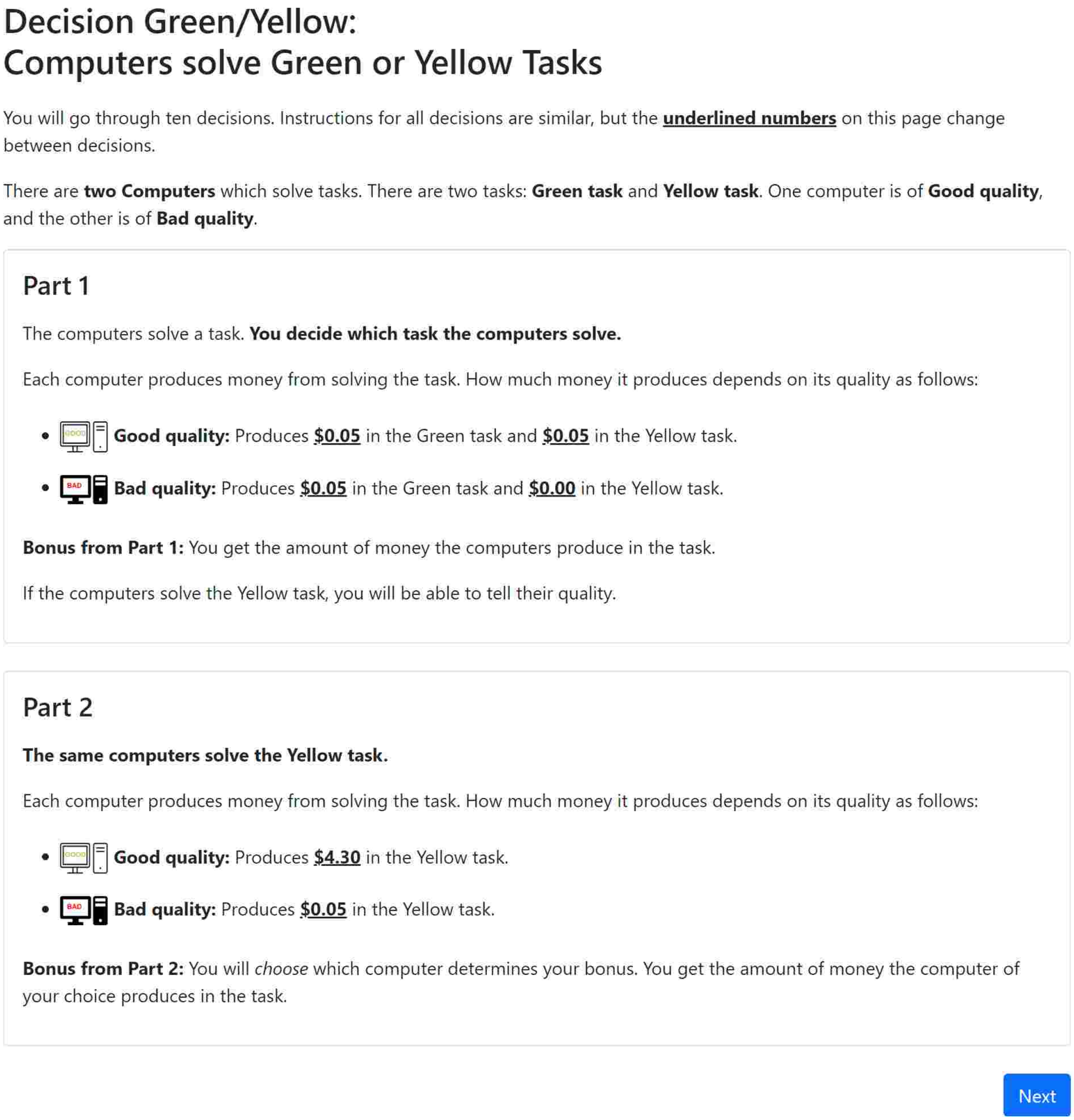
The same computers solve the Yellow task.

Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$4.30** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Yellow task.

Bonus from Part 2: You will *choose* which computer determines your bonus. You get the amount of money the computer of your choice produces in the task.

Figure A.9
Note: *Baseline* treatment.



Decision Green/Yellow:
Computers solve Green or Yellow Tasks

You will go through ten decisions. Instructions for all decisions are similar, but the **underlined numbers** on this page change between decisions.

There are **two Computers** which solve tasks. There are two tasks: **Green task** and **Yellow task**. One computer is of **Good quality**, and the other is of **Bad quality**.

Part 1

The computers solve a task. **You decide which task the computers solve.**

Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$0.05** in the Green task and **$0.05** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Green task and **$0.00** in the Yellow task.

Bonus from Part 1: You get the amount of money the computers produce in the task.

If the computers solve the Yellow task, you will be able to tell their quality.

Part 2

The same computers solve the Yellow task.

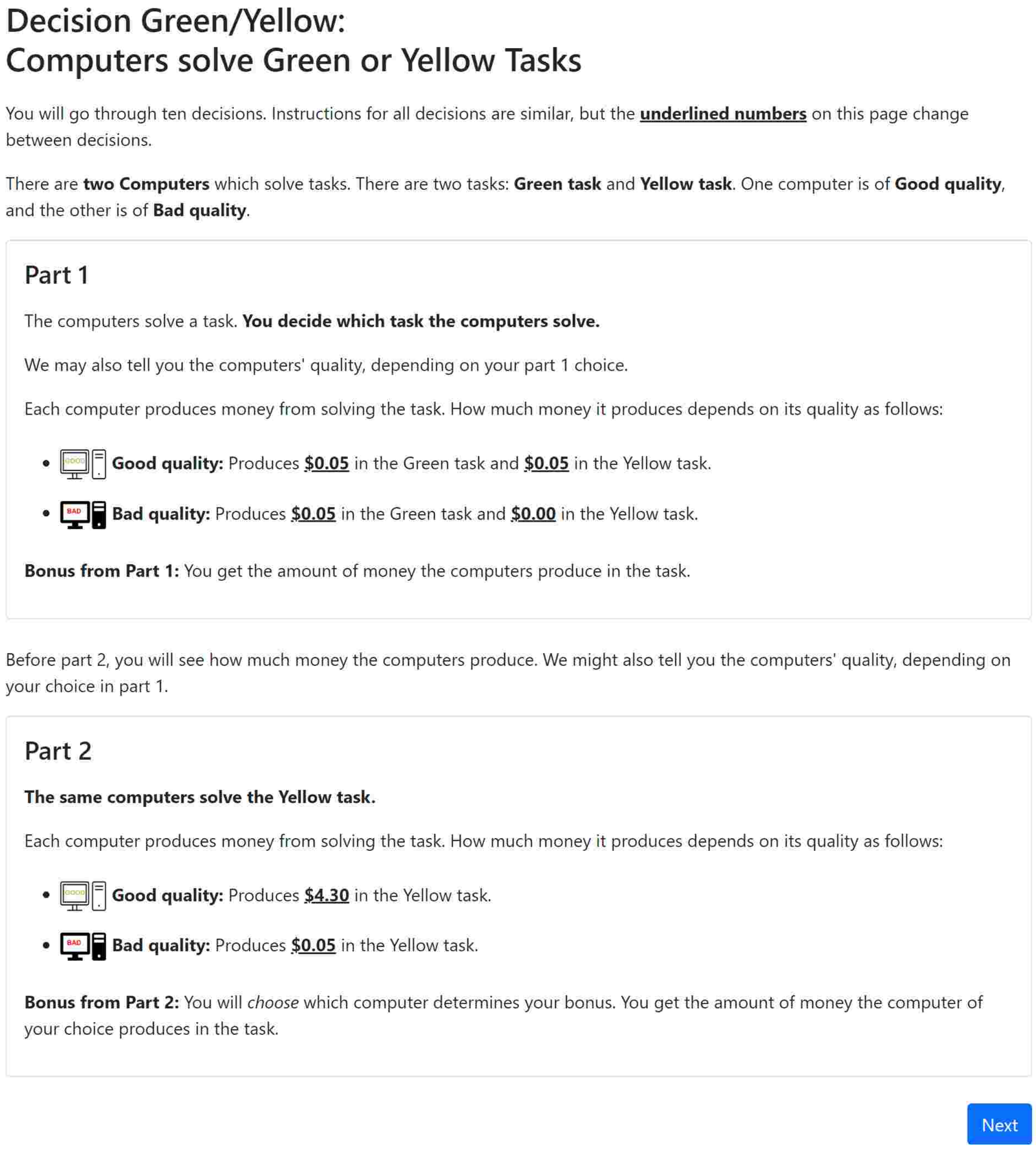
Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$4.30** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Yellow task.

Bonus from Part 2: You will *choose* which computer determines your bonus. You get the amount of money the computer of your choice produces in the task.

Figure A.10

Note: *Strategy Method* treatment.



Decision Green/Yellow:
Computers solve Green or Yellow Tasks

You will go through ten decisions. Instructions for all decisions are similar, but the **underlined numbers** on this page change between decisions.

There are **two Computers** which solve tasks. There are two tasks: **Green task** and **Yellow task**. One computer is of **Good quality**, and the other is of **Bad quality**.

Part 1

The computers solve a task. **You decide which task the computers solve.**

We may also tell you the computers' quality, depending on your part 1 choice.

Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$0.05** in the Green task and **$0.05** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Green task and **$0.00** in the Yellow task.

Bonus from Part 1: You get the amount of money the computers produce in the task.

Before part 2, you will see how much money the computers produce. We might also tell you the computers' quality, depending on your choice in part 1.

Part 2

The same computers solve the Yellow task.

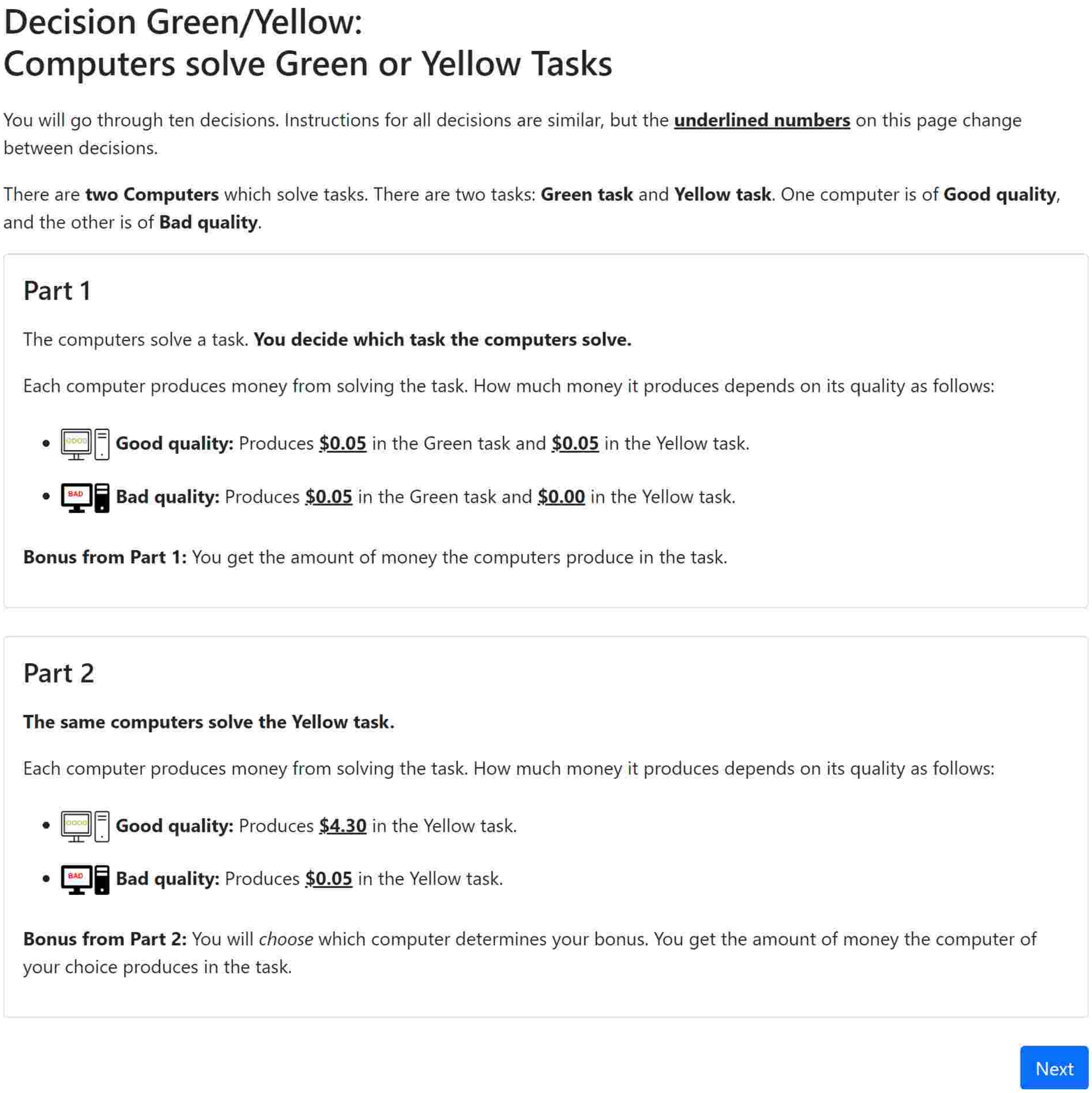
Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$4.30** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Yellow task.

Bonus from Part 2: You will *choose* which computer determines your bonus. You get the amount of money the computer of your choice produces in the task.

Figure A.11

Note: *Automatic Inference* treatment.



Decision Green/Yellow:
Computers solve Green or Yellow Tasks

You will go through ten decisions. Instructions for all decisions are similar, but the **underlined numbers** on this page change between decisions.

There are **two Computers** which solve tasks. There are two tasks: **Green task** and **Yellow task**. One computer is of **Good quality**, and the other is of **Bad quality**.

Part 1

The computers solve a task. **You decide which task the computers solve.**

Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$0.05** in the Green task and **$0.05** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Green task and **$0.00** in the Yellow task.

Bonus from Part 1: You get the amount of money the computers produce in the task.

Part 2

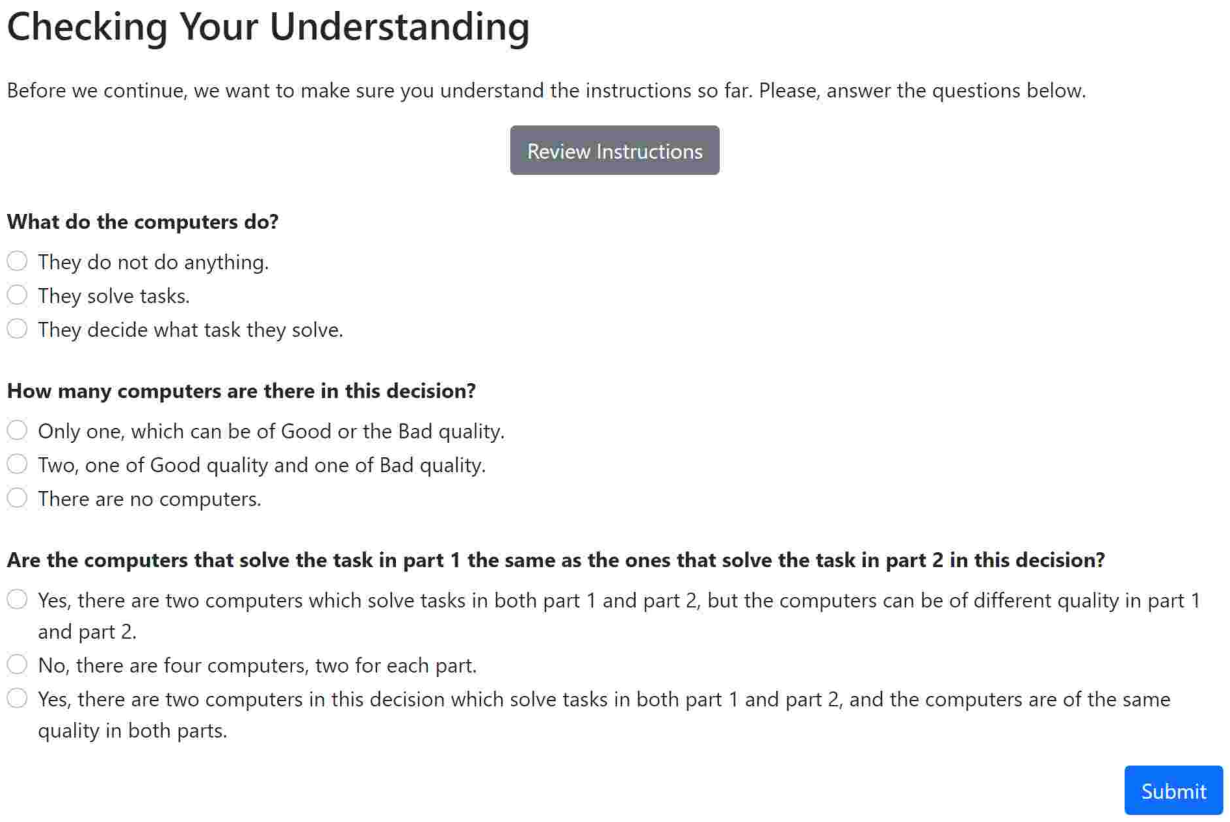
The same computers solve the Yellow task.

Each computer produces money from solving the task. How much money it produces depends on its quality as follows:

- **Good quality:** Produces **$4.30** in the Yellow task.
- **Bad quality:** Produces **$0.05** in the Yellow task.

Bonus from Part 2: You will *choose* which computer determines your bonus. You get the amount of money the computer of your choice produces in the task.

Figure A.12
Note: *Plan* treatment.



Checking Your Understanding

Before we continue, we want to make sure you understand the instructions so far. Please, answer the questions below.

What do the computers do?

○ They do not do anything.
○ They solve tasks.
○ They decide what task they solve.

How many computers are there in this decision?

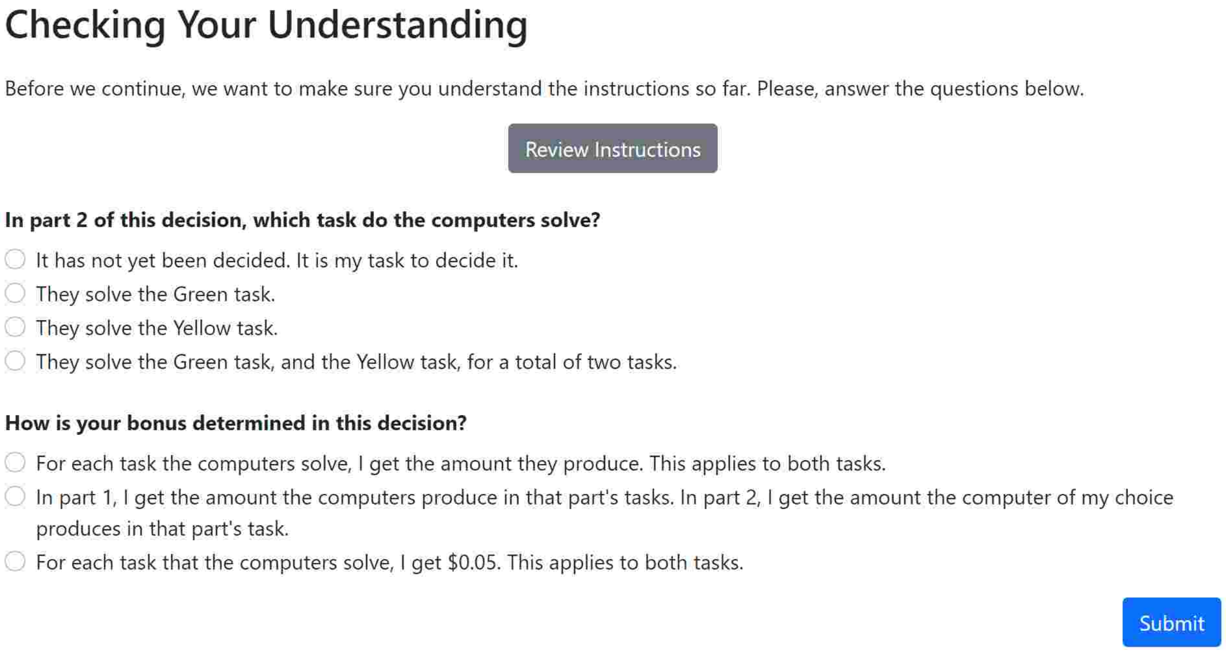
○ Only one, which can be of Good or the Bad quality.
○ Two, one of Good quality and one of Bad quality.
○ There are no computers.

Are the computers that solve the task in part 1 the same as the ones that solve the task in part 2 in this decision?

○ Yes, there are two computers which solve tasks in both part 1 and part 2, but the computers can be of different quality in part 1 and part 2.
○ No, there are four computers, two for each part.
○ Yes, there are two computers in this decision which solve tasks in both part 1 and part 2, and the computers are of the same quality in both parts.

Figure A.13

Note: All treatments.



Checking Your Understanding

Before we continue, we want to make sure you understand the instructions so far. Please, answer the questions below.

In part 2 of this decision, which task do the computers solve?

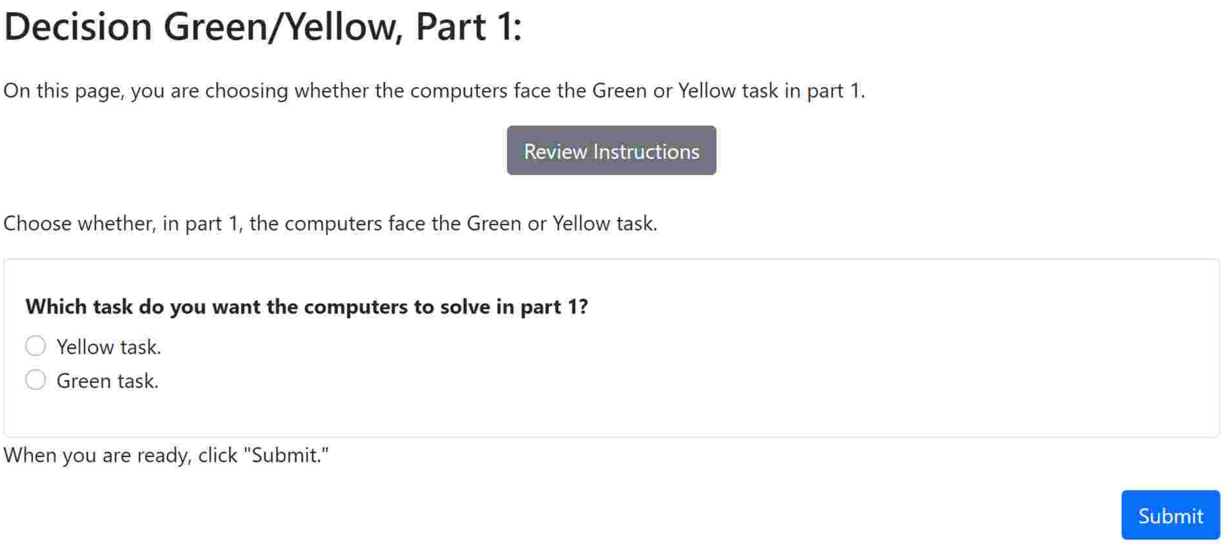
○ It has not yet been decided. It is my task to decide it.
○ They solve the Green task.
○ They solve the Yellow task.
○ They solve the Green task, and the Yellow task, for a total of two tasks.

How is your bonus determined in this decision?

○ For each task the computers solve, I get the amount they produce. This applies to both tasks.
○ In part 1, I get the amount the computers produce in that part's tasks. In part 2, I get the amount the computer of my choice produces in that part's task.
○ For each task that the computers solve, I get $0.05. This applies to both tasks.

Figure A.14
Note: All treatments.

Main decision: Baseline treatment



Decision Green/Yellow, Part 1:

On this page, you are choosing whether the computers face the Green or Yellow task in part 1.

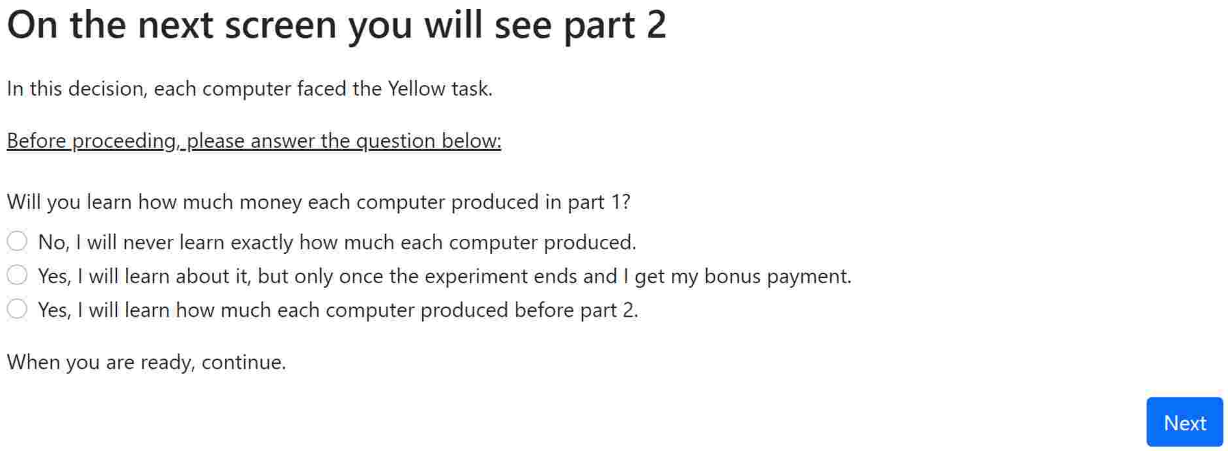
Choose whether, in part 1, the computers face the Green or Yellow task.

Which task do you want the computers to solve in part 1?

○ Yellow task.
○ Green task.

When you are ready, click "Submit."

Figure A.15
Note: *Baseline* treatment.



On the next screen you will see part 2

In this decision, each computer faced the Yellow task.

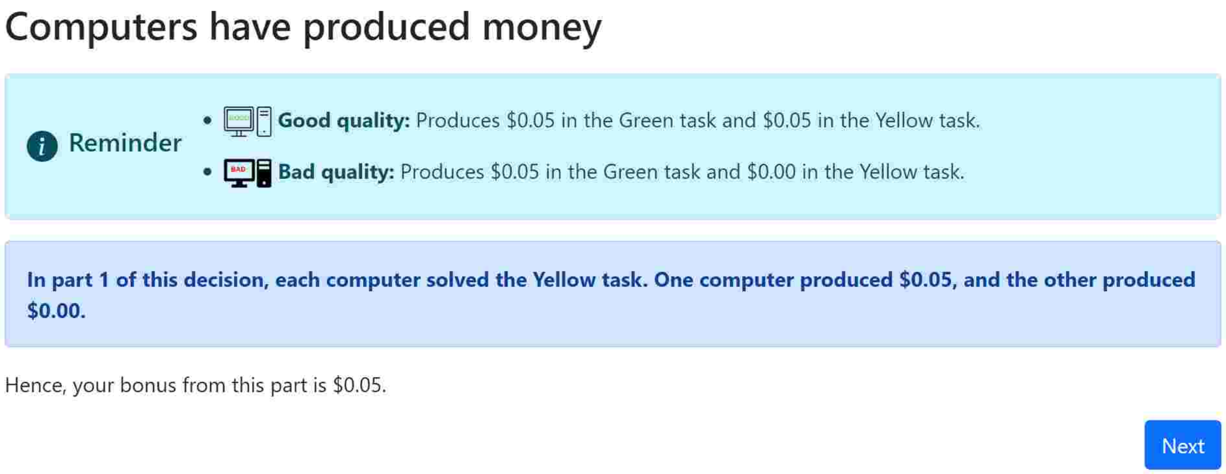
Before proceeding, please answer the question below:

Will you learn how much money each computer produced in part 1?

○ No, I will never learn exactly how much each computer produced.
○ Yes, I will learn about it, but only once the experiment ends and I get my bonus payment.
○ Yes, I will learn how much each computer produced before part 2.

When you are ready, continue.

Figure A.16
Note: *Baseline* treatment.



Computers have produced money

Reminder

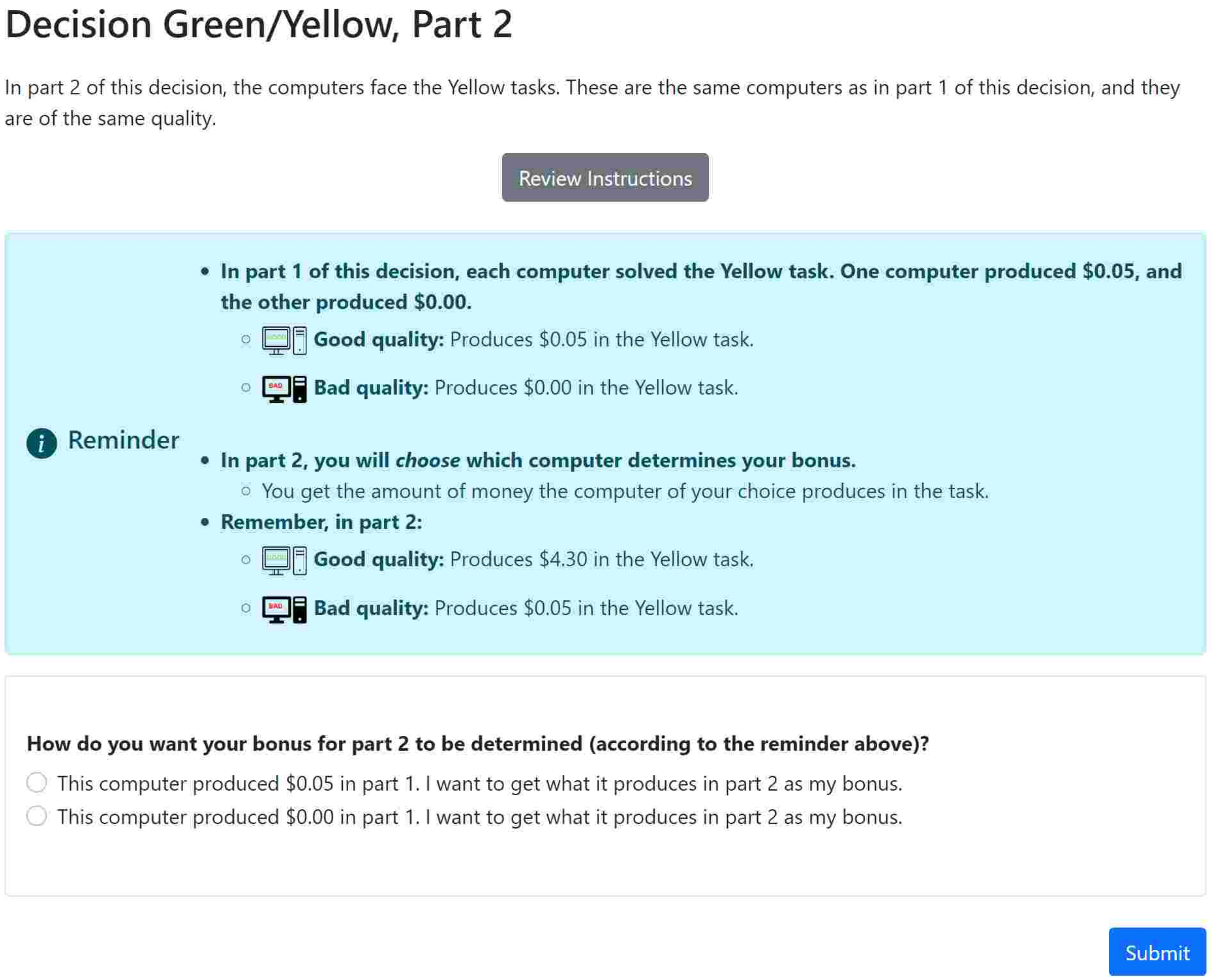
- **Good quality:** Produces $0.05 in the Green task and $0.05 in the Yellow task.
- **Bad quality:** Produces $0.05 in the Green task and $0.00 in the Yellow task.

In part 1 of this decision, each computer solved the Yellow task. One computer produced $0.05, and the other produced $0.00.

Hence, your bonus from this part is $0.05.

Figure A.17

Note: *Baseline* treatment.



Decision Green/Yellow, Part 2

In part 2 of this decision, the computers face the Yellow tasks. These are the same computers as in part 1 of this decision, and they are of the same quality.

Reminder

- **In part 1 of this decision, each computer solved the Yellow task. One computer produced $0.05, and the other produced $0.00.**
  - **Good quality:** Produces $0.05 in the Yellow task.
  - **Bad quality:** Produces $0.00 in the Yellow task.
- **In part 2, you will *choose* which computer determines your bonus.**
  - You get the amount of money the computer of your choice produces in the task.
- **Remember, in part 2:**
  - **Good quality:** Produces $4.30 in the Yellow task.
  - **Bad quality:** Produces $0.05 in the Yellow task.

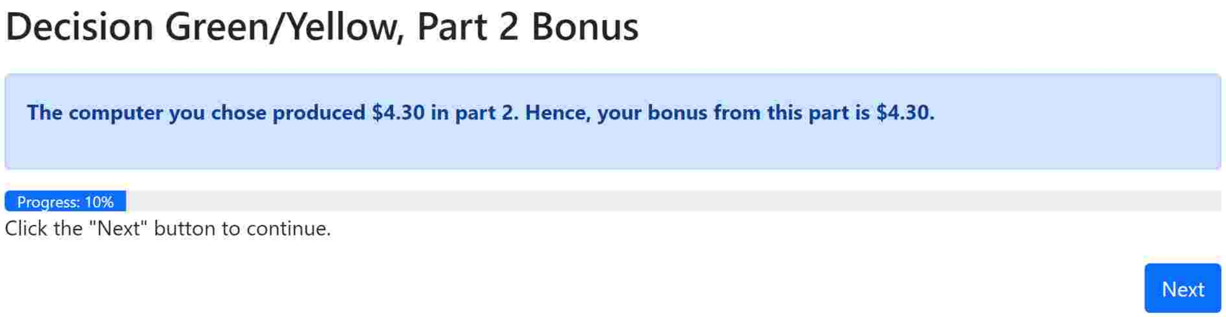
How do you want your bonus for part 2 to be determined (according to the reminder above)?

○ This computer produced $0.05 in part 1. I want to get what it produces in part 2 as my bonus.

○ This computer produced $0.00 in part 1. I want to get what it produces in part 2 as my bonus.

Figure A.18

Note: *Baseline* treatment.



Decision Green/Yellow, Part 2 Bonus

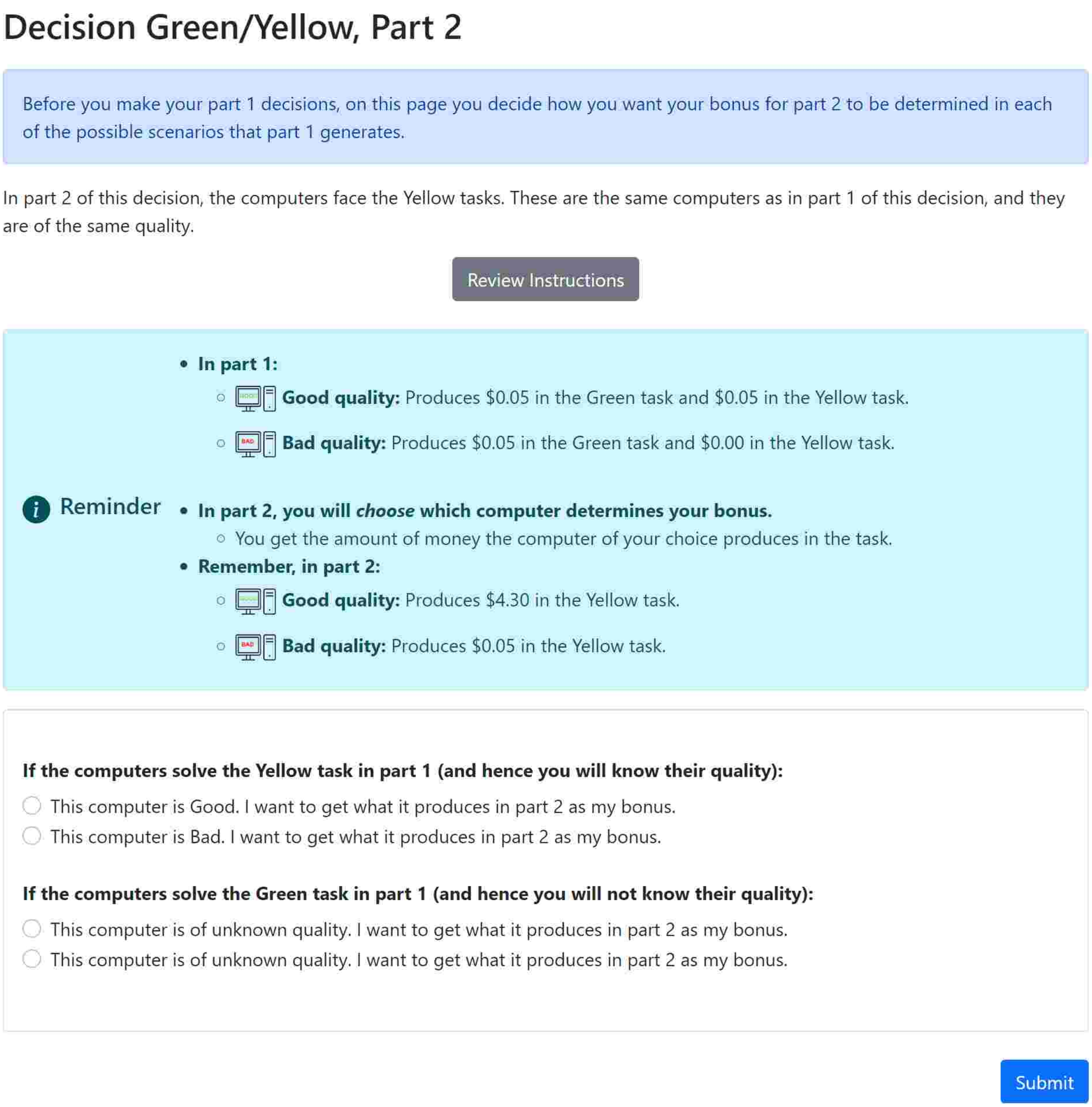
The computer you chose produced $4.30 in part 2. Hence, your bonus from this part is $4.30.

Click the "Next" button to continue.

Figure A.19

Note: *Baseline* treatment.

Main decision: Strategy Method treatment



Decision Green/Yellow, Part 2

Before you make your part 1 decisions, on this page you decide how you want your bonus for part 2 to be determined in each of the possible scenarios that part 1 generates.

In part 2 of this decision, the computers face the Yellow tasks. These are the same computers as in part 1 of this decision, and they are of the same quality.

Reminder

- **In part 1:**
  - **Good quality:** Produces $0.05 in the Green task and $0.05 in the Yellow task.
  - **Bad quality:** Produces $0.05 in the Green task and $0.00 in the Yellow task.
- **In part 2, you will *choose* which computer determines your bonus.**
  - You get the amount of money the computer of your choice produces in the task.
- **Remember, in part 2:**
  - **Good quality:** Produces $4.30 in the Yellow task.
  - **Bad quality:** Produces $0.05 in the Yellow task.

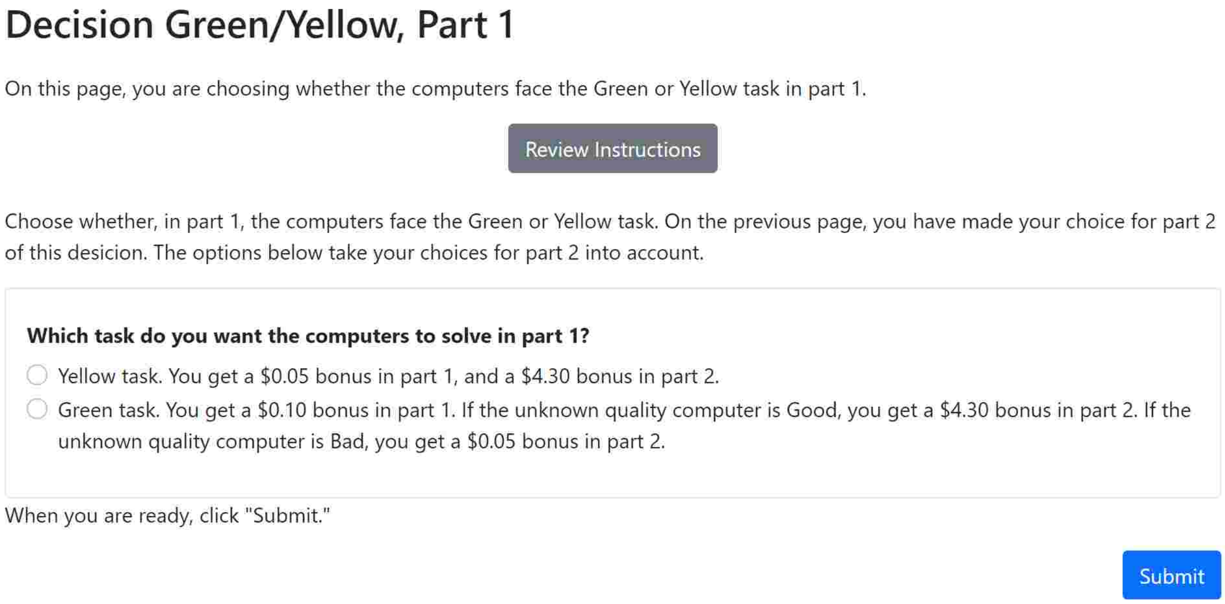
If the computers solve the Yellow task in part 1 (and hence you will know their quality):

○ This computer is Good. I want to get what it produces in part 2 as my bonus.
○ This computer is Bad. I want to get what it produces in part 2 as my bonus.

If the computers solve the Green task in part 1 (and hence you will not know their quality):

○ This computer is of unknown quality. I want to get what it produces in part 2 as my bonus.
○ This computer is of unknown quality. I want to get what it produces in part 2 as my bonus.

Figure A.20
Note: *Strategy Method* treatment.



Decision Green/Yellow, Part 1

On this page, you are choosing whether the computers face the Green or Yellow task in part 1.

Choose whether, in part 1, the computers face the Green or Yellow task. On the previous page, you have made your choice for part 2 of this desicion. The options below take your choices for part 2 into account.

Which task do you want the computers to solve in part 1?

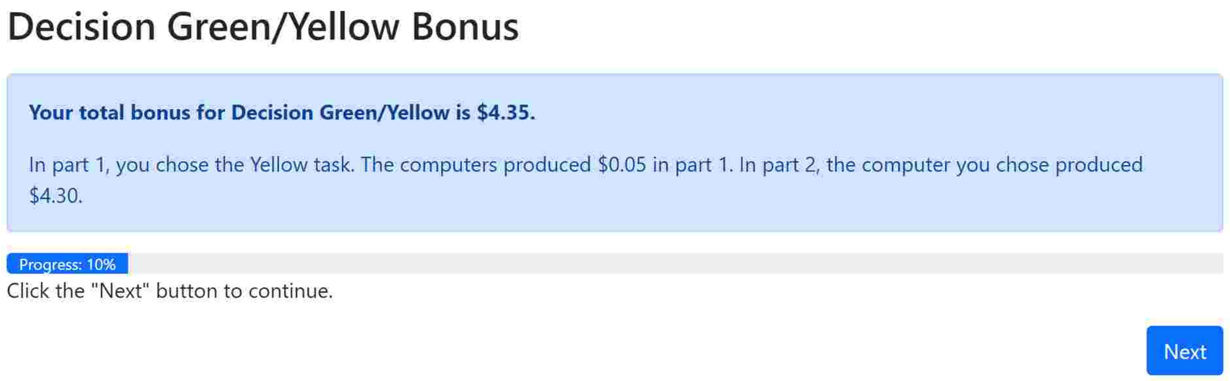
○ Yellow task. You get a $0.05 bonus in part 1, and a $4.30 bonus in part 2.

○ Green task. You get a $0.10 bonus in part 1. If the unknown quality computer is Good, you get a $4.30 bonus in part 2. If the unknown quality computer is Bad, you get a $0.05 bonus in part 2.

When you are ready, click "Submit."

Figure A.21

Note: *Strategy Method* treatment.



Decision Green/Yellow Bonus

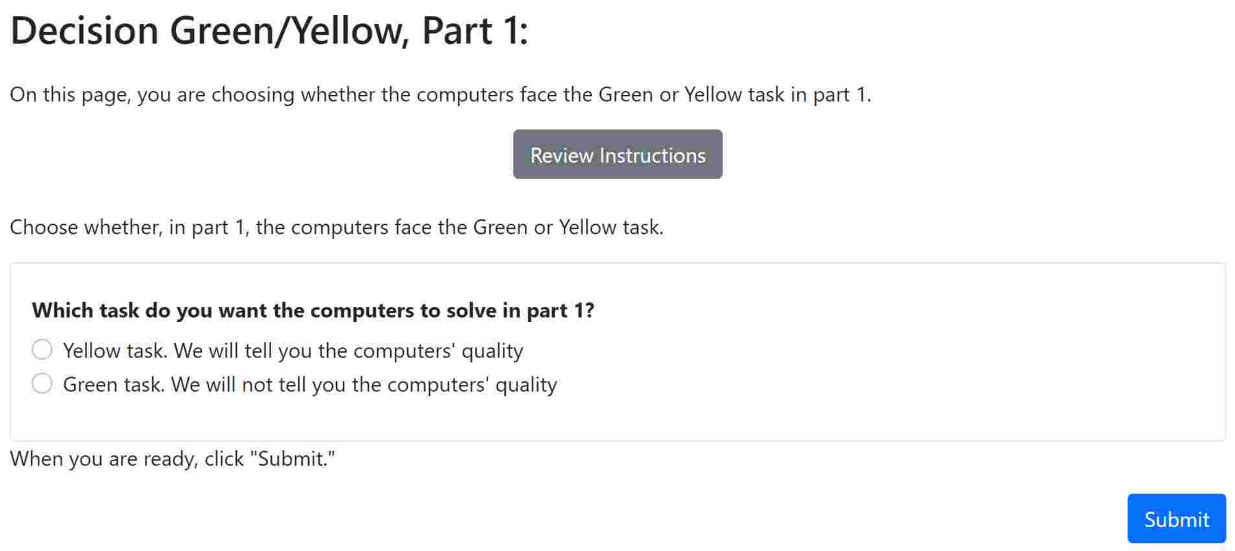
Your total bonus for Decision Green/Yellow is $4.35.

In part 1, you chose the Yellow task. The computers produced $0.05 in part 1. In part 2, the computer you chose produced $4.30.

Click the "Next" button to continue.

Figure A.22
Note: *Strategy Method* treatment.

Main decision: Automatic Inference treatment



Decision Green/Yellow, Part 1:

On this page, you are choosing whether the computers face the Green or Yellow task in part 1.

Choose whether, in part 1, the computers face the Green or Yellow task.

Which task do you want the computers to solve in part 1?

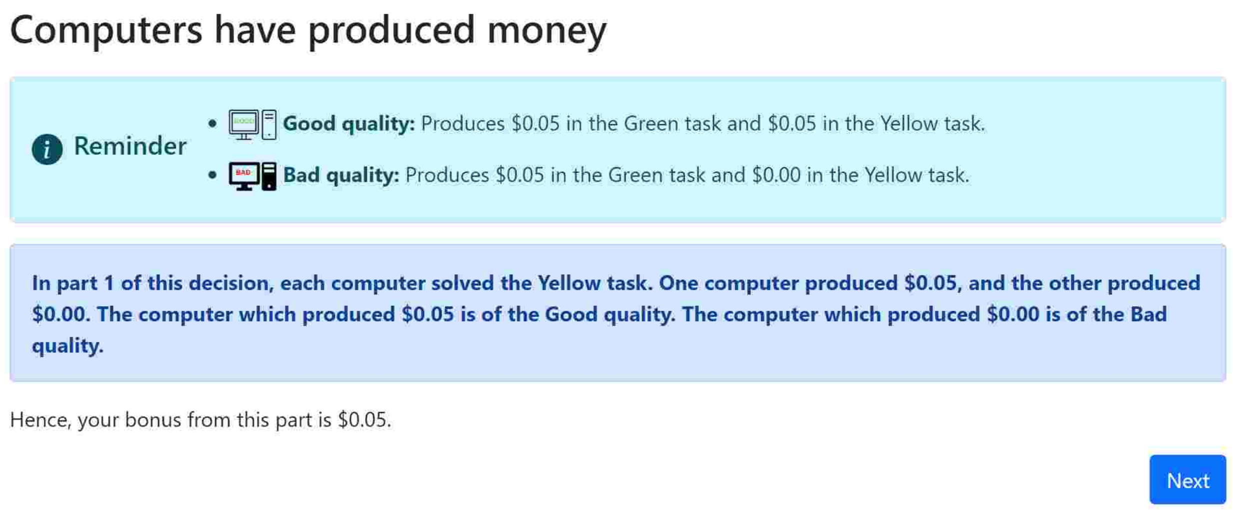
○ Yellow task. We will tell you the computers' quality

○ Green task. We will not tell you the computers' quality

When you are ready, click "Submit."

Figure A.23

Note: *Automatic Inference* treatment.



Computers have produced money

Reminder

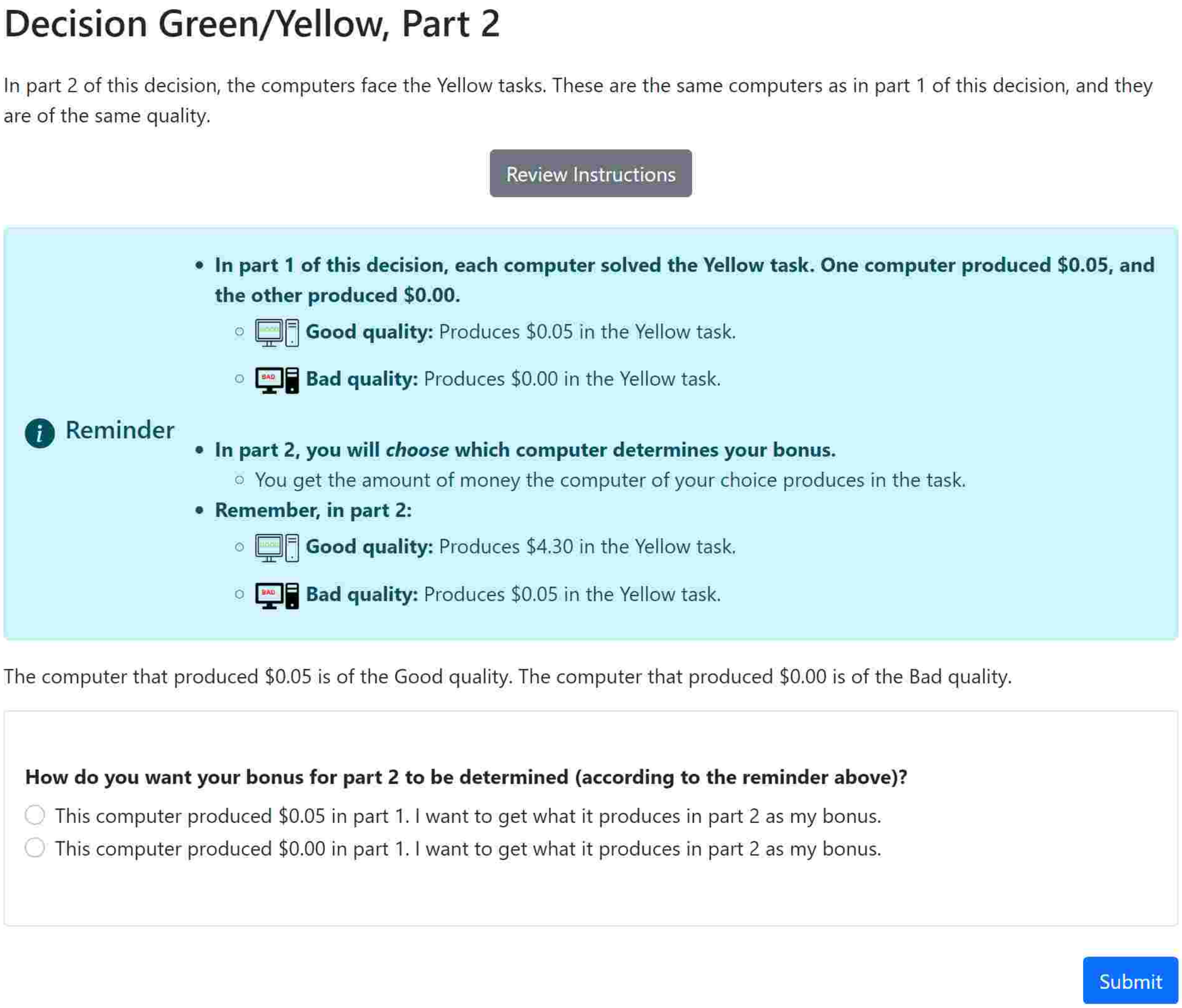
- Good quality: Produces $0.05 in the Green task and $0.05 in the Yellow task.
- Bad quality: Produces $0.05 in the Green task and $0.00 in the Yellow task.

In part 1 of this decision, each computer solved the Yellow task. One computer produced $0.05, and the other produced $0.00. The computer which produced $0.05 is of the Good quality. The computer which produced $0.00 is of the Bad quality.

Hence, your bonus from this part is $0.05.

Figure A.24

Note: *Automatic Inference* treatment.



Decision Green/Yellow, Part 2

In part 2 of this decision, the computers face the Yellow tasks. These are the same computers as in part 1 of this decision, and they are of the same quality.

- **In part 1 of this decision, each computer solved the Yellow task. One computer produced $0.05, and the other produced $0.00.**
  - **Good quality:** Produces $0.05 in the Yellow task.
  - **Bad quality:** Produces $0.00 in the Yellow task.
- **In part 2, you will *choose* which computer determines your bonus.**
  - You get the amount of money the computer of your choice produces in the task.
- **Remember, in part 2:**
  - **Good quality:** Produces $4.30 in the Yellow task.
  - **Bad quality:** Produces $0.05 in the Yellow task.

The computer that produced $0.05 is of the Good quality. The computer that produced $0.00 is of the Bad quality.

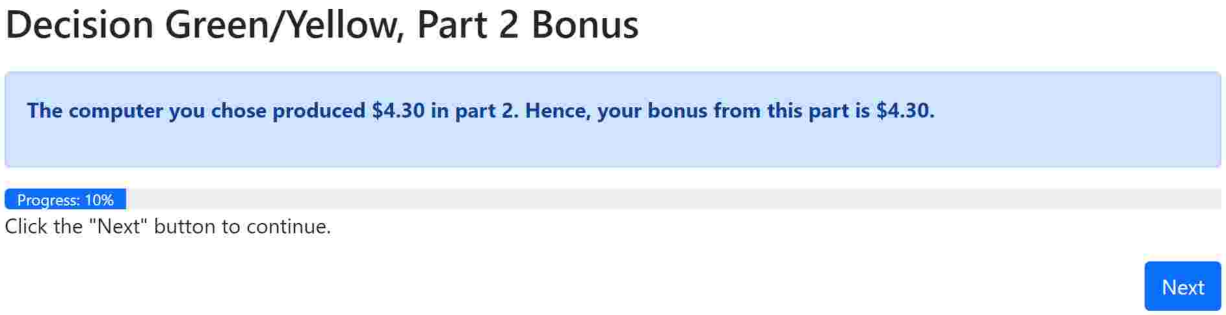
How do you want your bonus for part 2 to be determined (according to the reminder above)?

○ This computer produced $0.05 in part 1. I want to get what it produces in part 2 as my bonus.

○ This computer produced $0.00 in part 1. I want to get what it produces in part 2 as my bonus.

Figure A.25

Note: *Automatic Inference* treatment.



Decision Green/Yellow, Part 2 Bonus

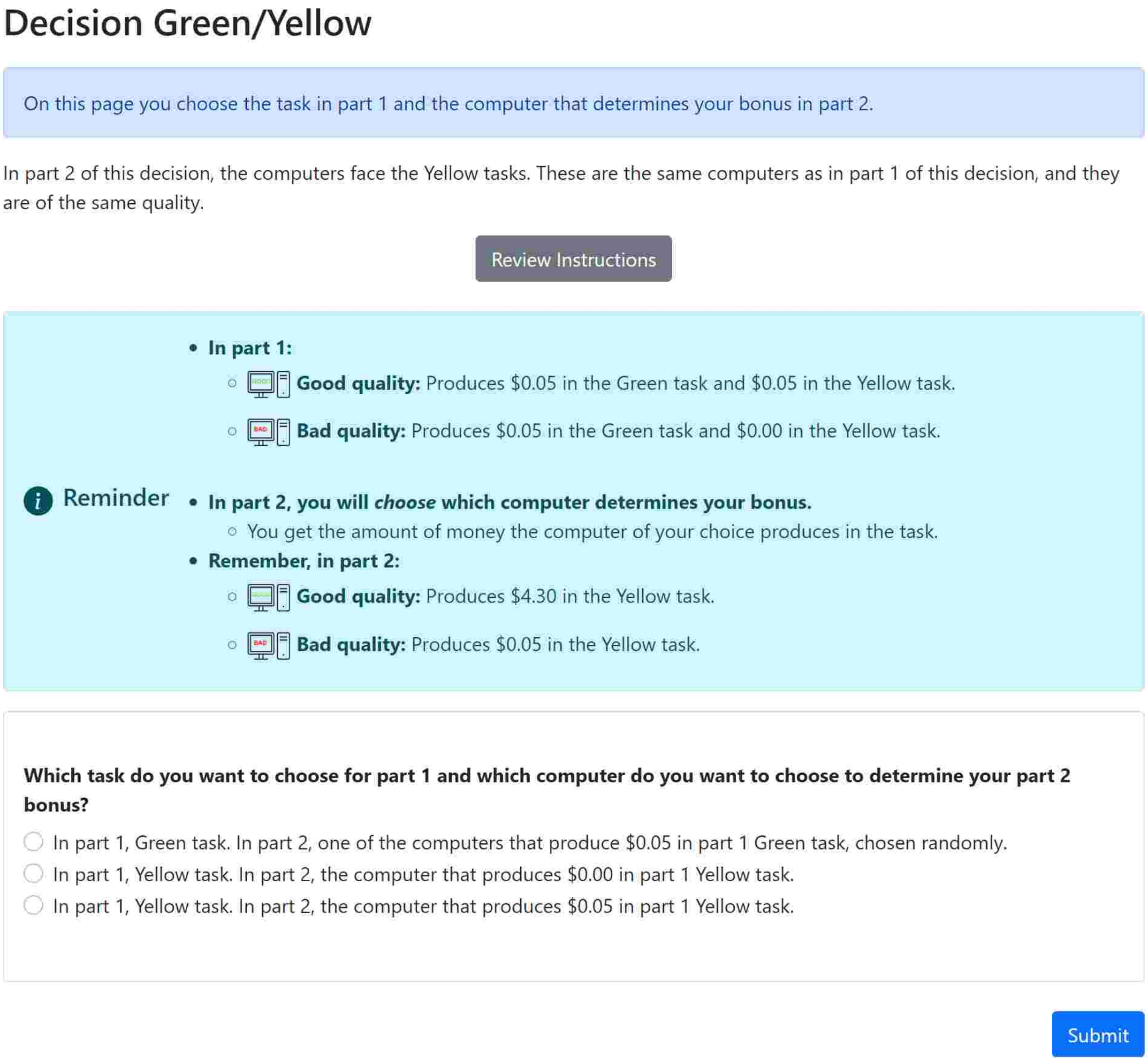
The computer you chose produced $4.30 in part 2. Hence, your bonus from this part is $4.30.

Click the "Next" button to continue.

Figure A.26

Note: *Automatic Inference* treatment.

Main decision: Plan treatment



Decision Green/Yellow

On this page you choose the task in part 1 and the computer that determines your bonus in part 2.

In part 2 of this decision, the computers face the Yellow tasks. These are the same computers as in part 1 of this decision, and they are of the same quality.

Reminder

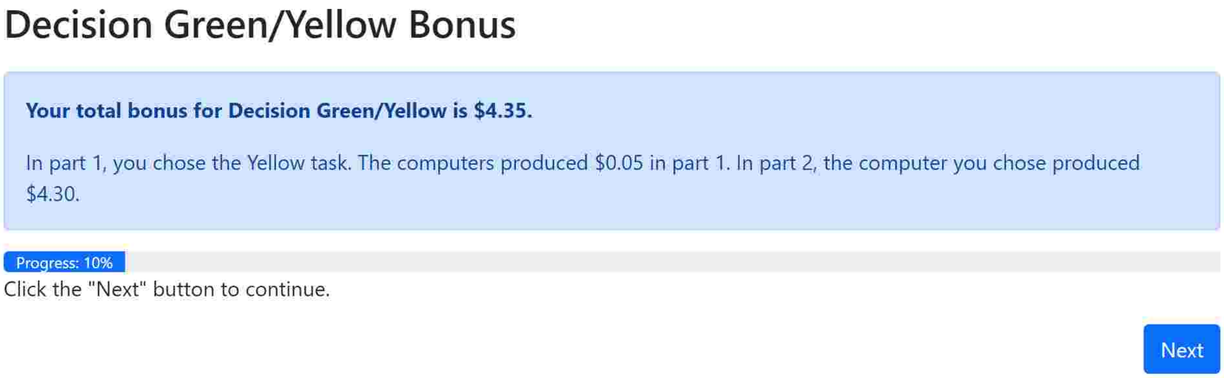
- **In part 1:**
  - **Good quality:** Produces $0.05 in the Green task and $0.05 in the Yellow task.
  - **Bad quality:** Produces $0.05 in the Green task and $0.00 in the Yellow task.
- **In part 2, you will *choose* which computer determines your bonus.**
  - You get the amount of money the computer of your choice produces in the task.
- **Remember, in part 2:**
  - **Good quality:** Produces $4.30 in the Yellow task.
  - **Bad quality:** Produces $0.05 in the Yellow task.

Which task do you want to choose for part 1 and which computer do you want to choose to determine your part 2 bonus?

○ In part 1, Green task. In part 2, one of the computers that produce $0.05 in part 1 Green task, chosen randomly.
○ In part 1, Yellow task. In part 2, the computer that produces $0.00 in part 1 Yellow task.
○ In part 1, Yellow task. In part 2, the computer that produces $0.05 in part 1 Yellow task.

Figure A.27

Note: *Plan* treatment.



Decision Green/Yellow Bonus

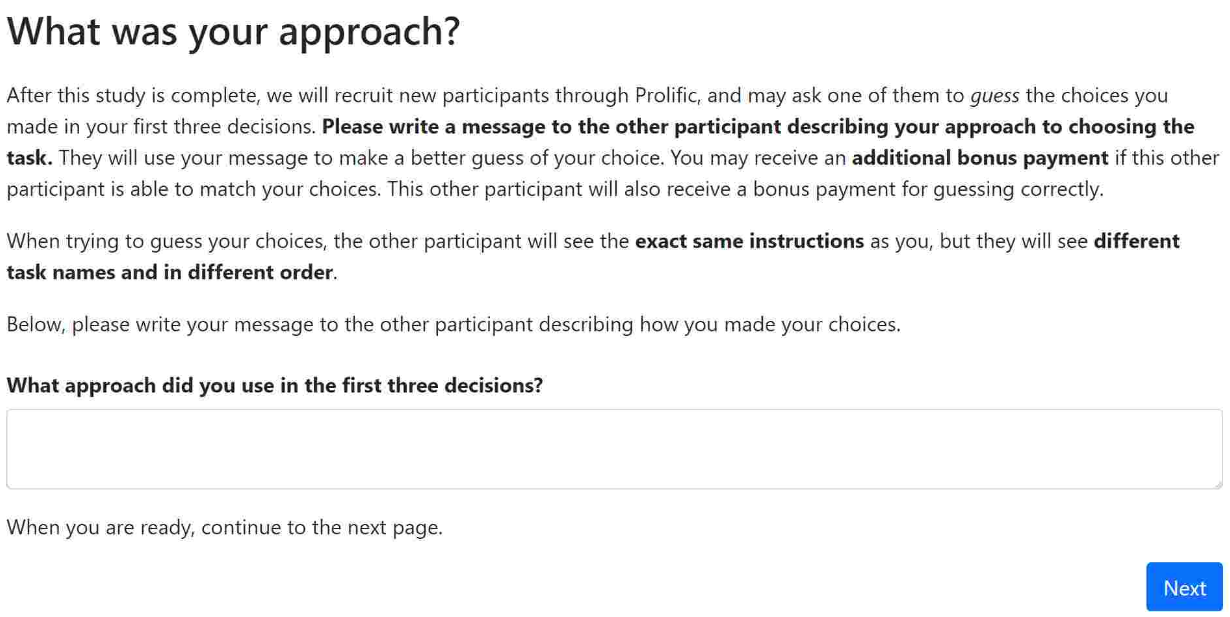
Your total bonus for Decision Green/Yellow is $4.35.

In part 1, you chose the Yellow task. The computers produced $0.05 in part 1. In part 2, the computer you chose produced $4.30.

Click the "Next" button to continue.

Figure A.28
Note: *Plan* treatment.

Other elicitations



What was your approach?

After this study is complete, we will recruit new participants through Prolific, and may ask one of them to *guess* the choices you made in your first three decisions. **Please write a message to the other participant describing your approach to choosing the task.** They will use your message to make a better guess of your choice. You may receive an **additional bonus payment** if this other participant is able to match your choices. This other participant will also receive a bonus payment for guessing correctly.

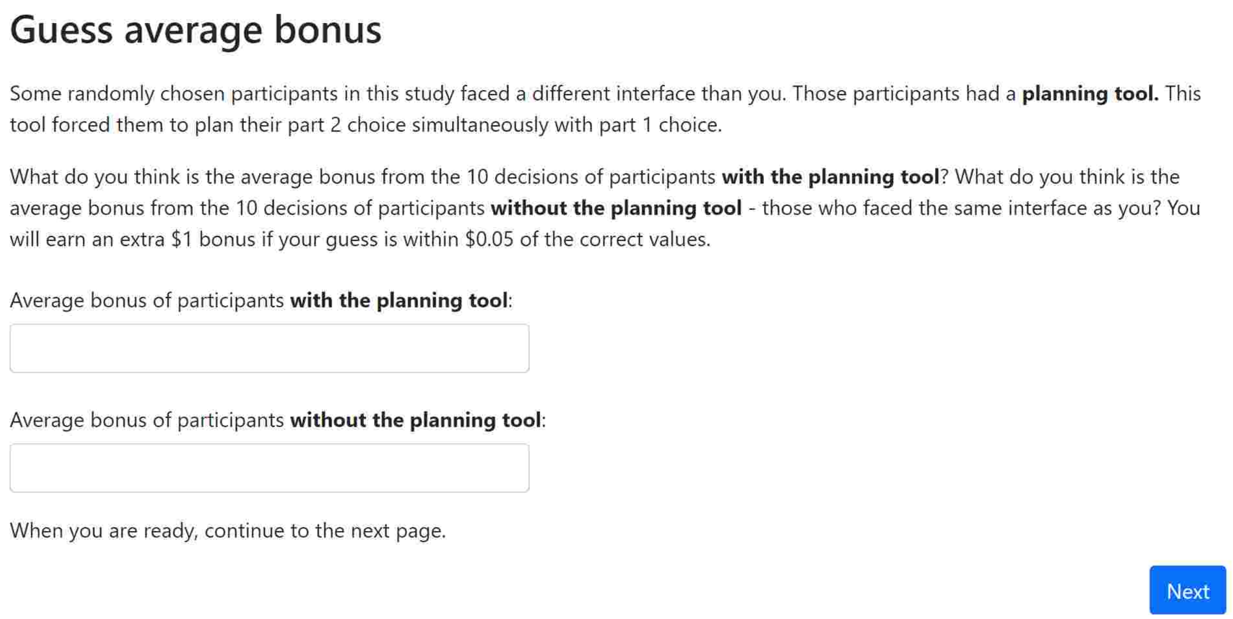
When trying to guess your choices, the other participant will see the **exact same instructions** as you, but they will see **different task names and in different order**.

Below, please write your message to the other participant describing how you made your choices.

What approach did you use in the first three decisions?

When you are ready, continue to the next page.

Figure A.29
<u>Note:</u> All treatments.



Guess average bonus

Some randomly chosen participants in this study faced a different interface than you. Those participants had a **planning tool.** This tool forced them to plan their part 2 choice simultaneously with part 1 choice.

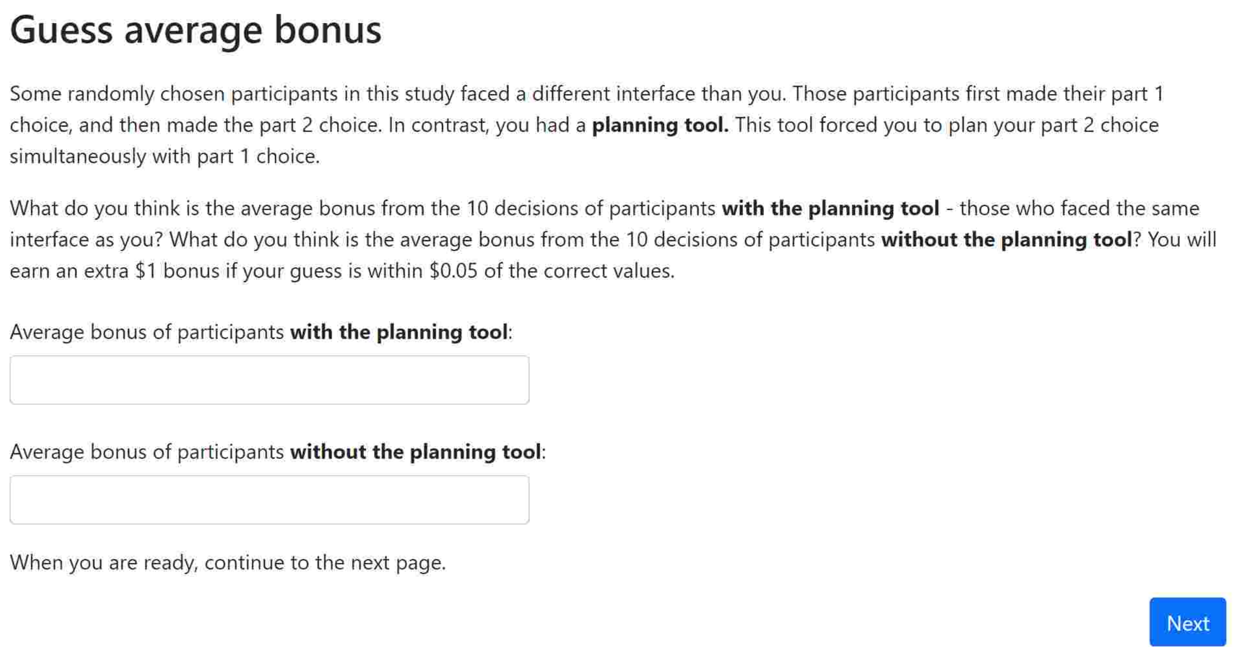
What do you think is the average bonus from the 10 decisions of participants **with the planning tool**? What do you think is the average bonus from the 10 decisions of participants **without the planning tool** - those who faced the same interface as you? You will earn an extra $1 bonus if your guess is within $0.05 of the correct values.

Average bonus of participants **with the planning tool**:

Average bonus of participants **without the planning tool**:

When you are ready, continue to the next page.

Figure A.30
Note: *Baseline* treatment.



Guess average bonus

Some randomly chosen participants in this study faced a different interface than you. Those participants first made their part 1 choice, and then made the part 2 choice. In contrast, you had a **planning tool.** This tool forced you to plan your part 2 choice simultaneously with part 1 choice.

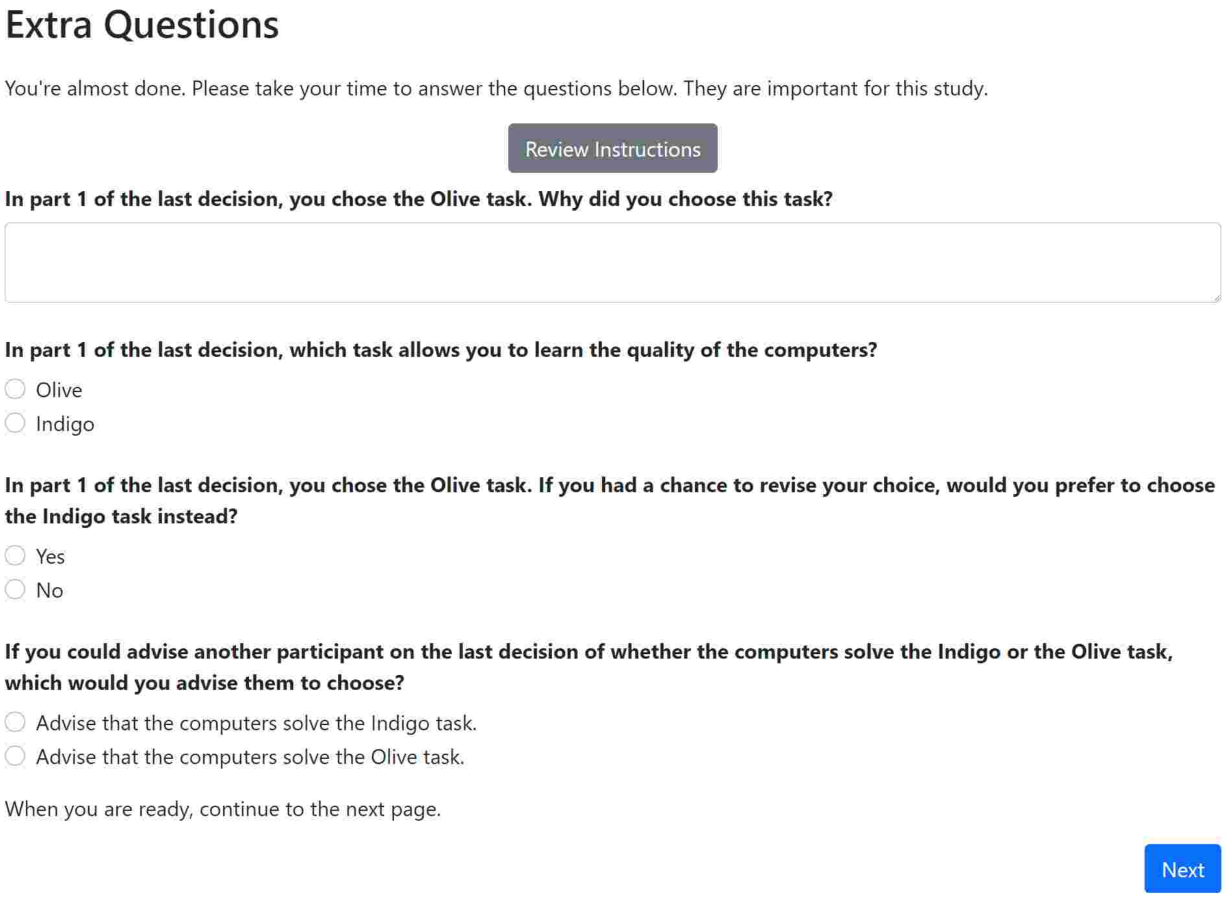
What do you think is the average bonus from the 10 decisions of participants **with the planning tool** - those who faced the same interface as you? What do you think is the average bonus from the 10 decisions of participants **without the planning tool**? You will earn an extra $1 bonus if your guess is within $0.05 of the correct values.

Average bonus of participants **with the planning tool**:

Average bonus of participants **without the planning tool**:

When you are ready, continue to the next page.

Figure A.31
Note: *Plan* treatment.



Extra Questions

You're almost done. Please take your time to answer the questions below. They are important for this study.

In part 1 of the last decision, you chose the Olive task. Why did you choose this task?

In part 1 of the last decision, which task allows you to learn the quality of the computers?

○ Olive
○ Indigo

In part 1 of the last decision, you chose the Olive task. If you had a chance to revise your choice, would you prefer to choose the Indigo task instead?

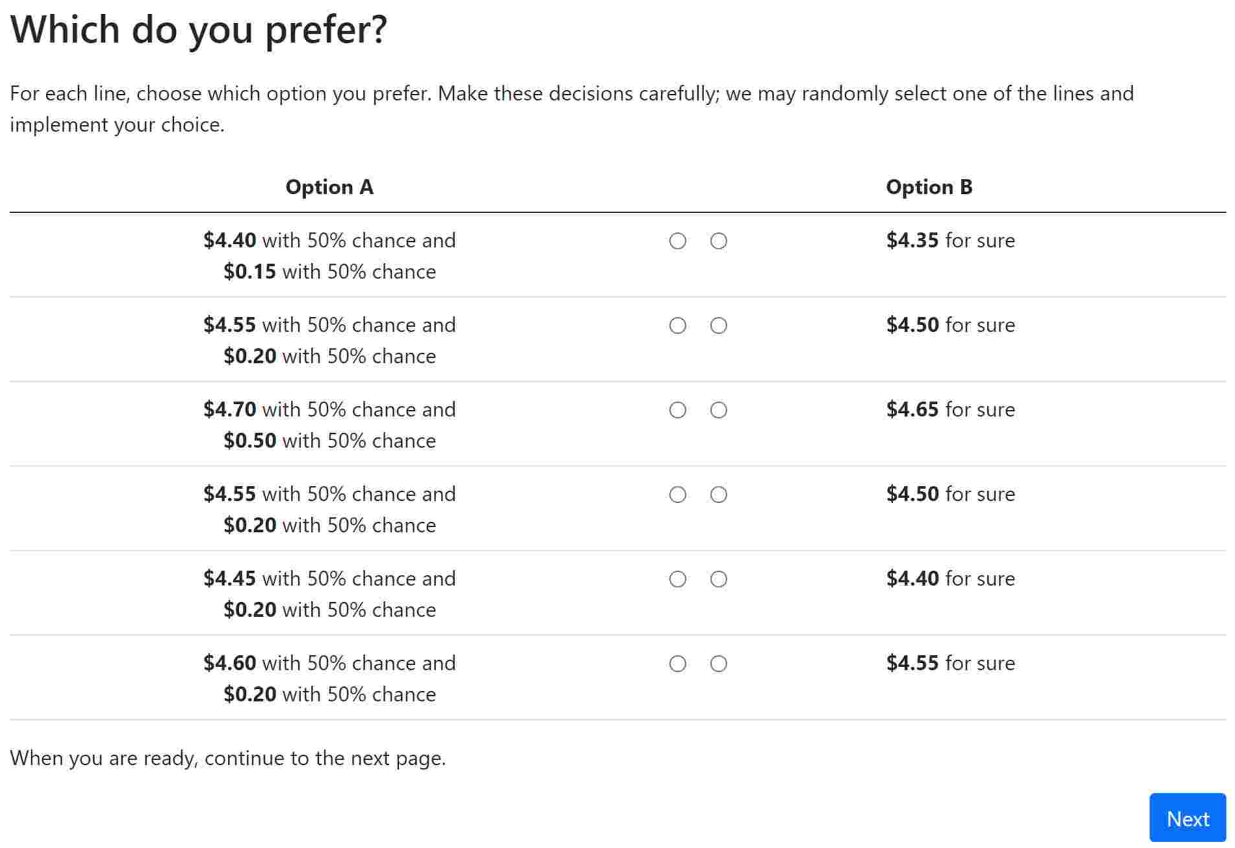
○ Yes
○ No

If you could advise another participant on the last decision of whether the computers solve the Indigo or the Olive task, which would you advise them to choose?

○ Advise that the computers solve the Indigo task.
○ Advise that the computers solve the Olive task.

When you are ready, continue to the next page.

Figure A.32
Note: All treatments.



Which do you prefer?

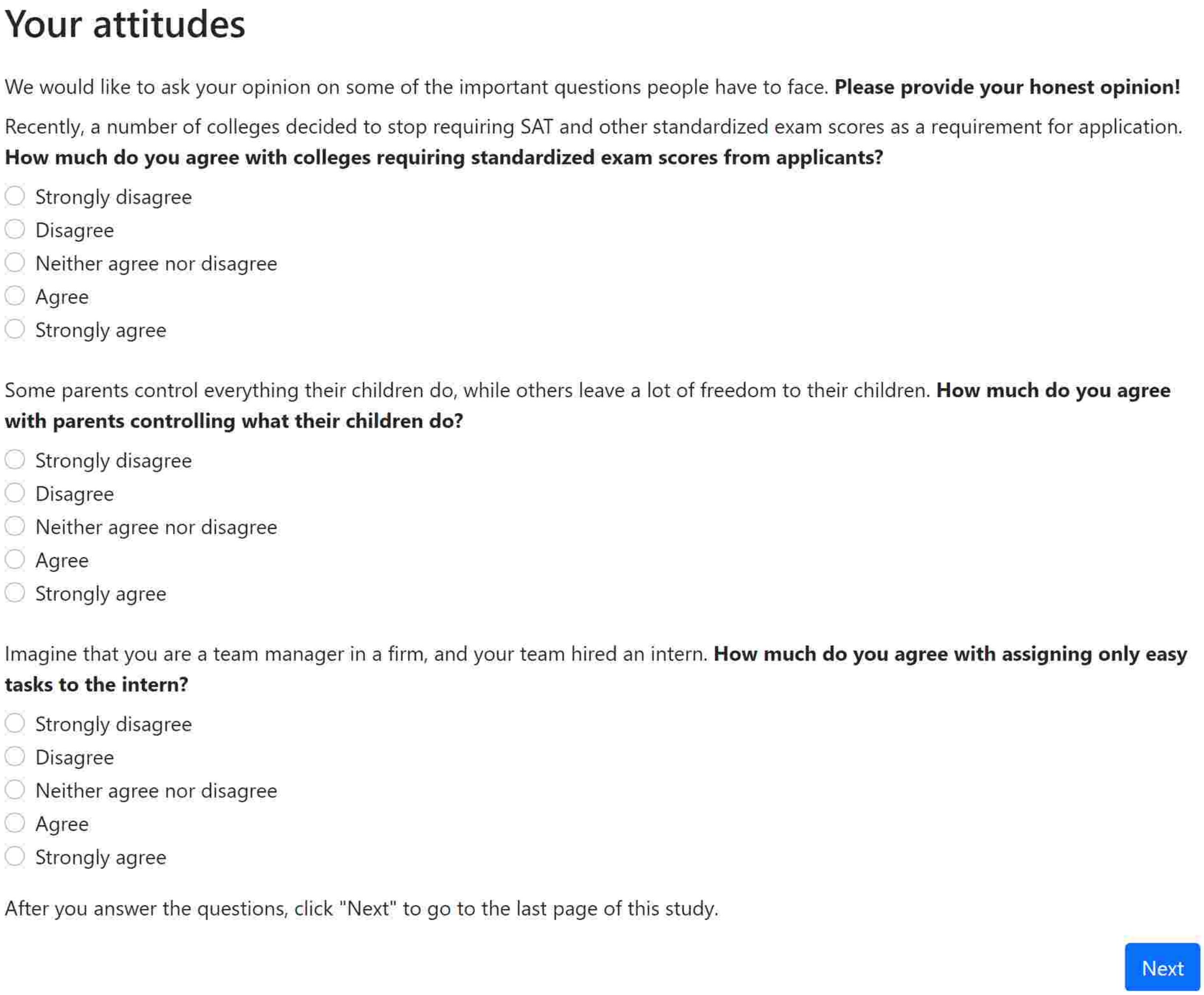
For each line, choose which option you prefer. Make these decisions carefully; we may randomly select one of the lines and implement your choice.

Option A		Option B
$4.40 with 50% chance and **$0.15** with 50% chance	○ ○	**$4.35** for sure
$4.55 with 50% chance and **$0.20** with 50% chance	○ ○	**$4.50** for sure
$4.70 with 50% chance and **$0.50** with 50% chance	○ ○	**$4.65** for sure
$4.55 with 50% chance and **$0.20** with 50% chance	○ ○	**$4.50** for sure
$4.45 with 50% chance and **$0.20** with 50% chance	○ ○	**$4.40** for sure
$4.60 with 50% chance and **$0.20** with 50% chance	○ ○	**$4.55** for sure

When you are ready, continue to the next page.

Figure A.33

Note: All treatments.



Your attitudes

We would like to ask your opinion on some of the important questions people have to face. **Please provide your honest opinion!**

Recently, a number of colleges decided to stop requiring SAT and other standardized exam scores as a requirement for application. **How much do you agree with colleges requiring standardized exam scores from applicants?**

○ Strongly disagree
○ Disagree
○ Neither agree nor disagree
○ Agree
○ Strongly agree

Some parents control everything their children do, while others leave a lot of freedom to their children. **How much do you agree with parents controlling what their children do?**

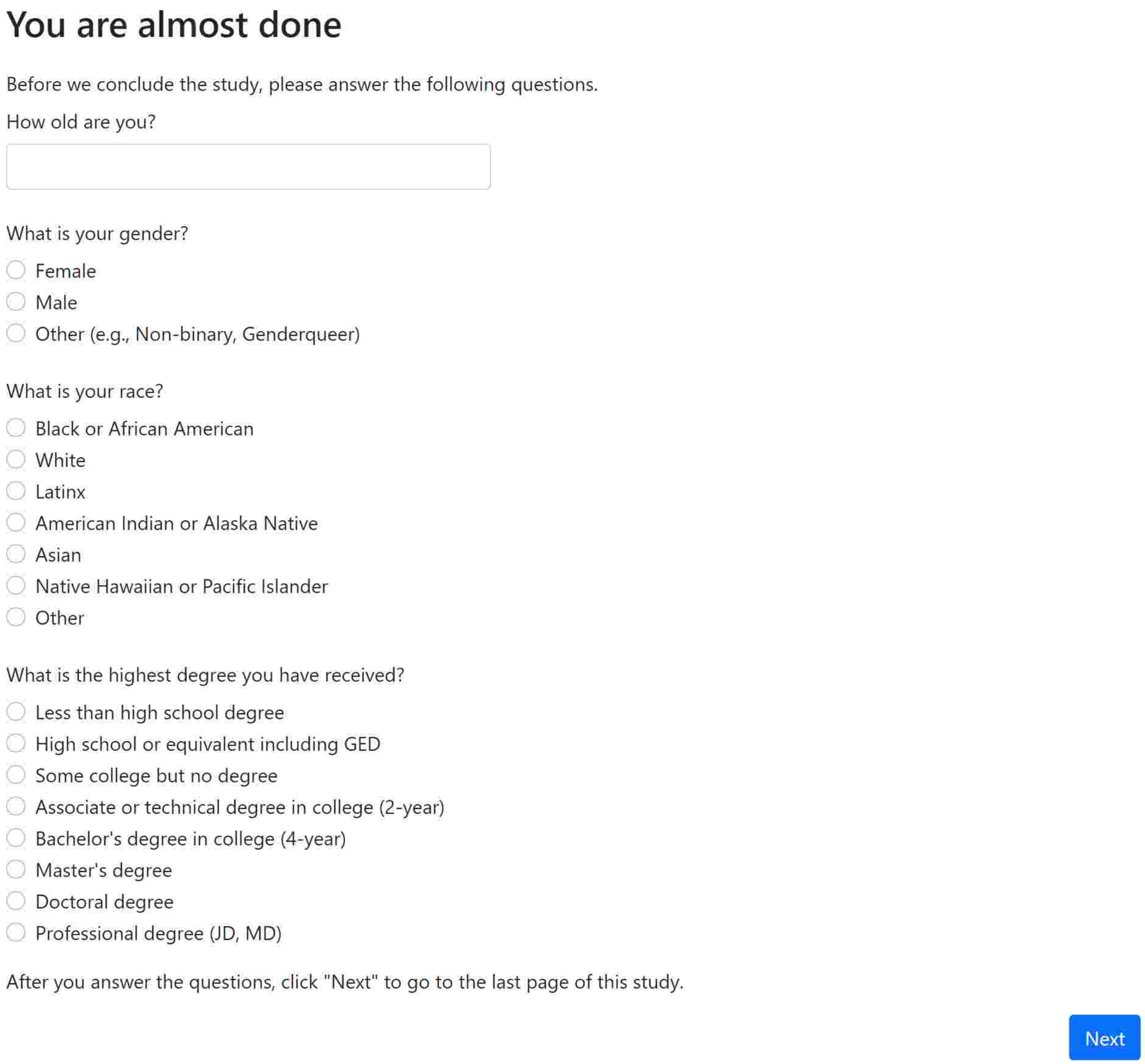
○ Strongly disagree
○ Disagree
○ Neither agree nor disagree
○ Agree
○ Strongly agree

Imagine that you are a team manager in a firm, and your team hired an intern. **How much do you agree with assigning only easy tasks to the intern?**

○ Strongly disagree
○ Disagree
○ Neither agree nor disagree
○ Agree
○ Strongly agree

After you answer the questions, click "Next" to go to the last page of this study.

Figure A.34
Note: All treatments.



You are almost done

Before we conclude the study, please answer the following questions.

How old are you?

What is your gender?

○ Female
○ Male
○ Other (e.g., Non-binary, Genderqueer)

What is your race?

○ Black or African American
○ White
○ Latinx
○ American Indian or Alaska Native
○ Asian
○ Native Hawaiian or Pacific Islander
○ Other

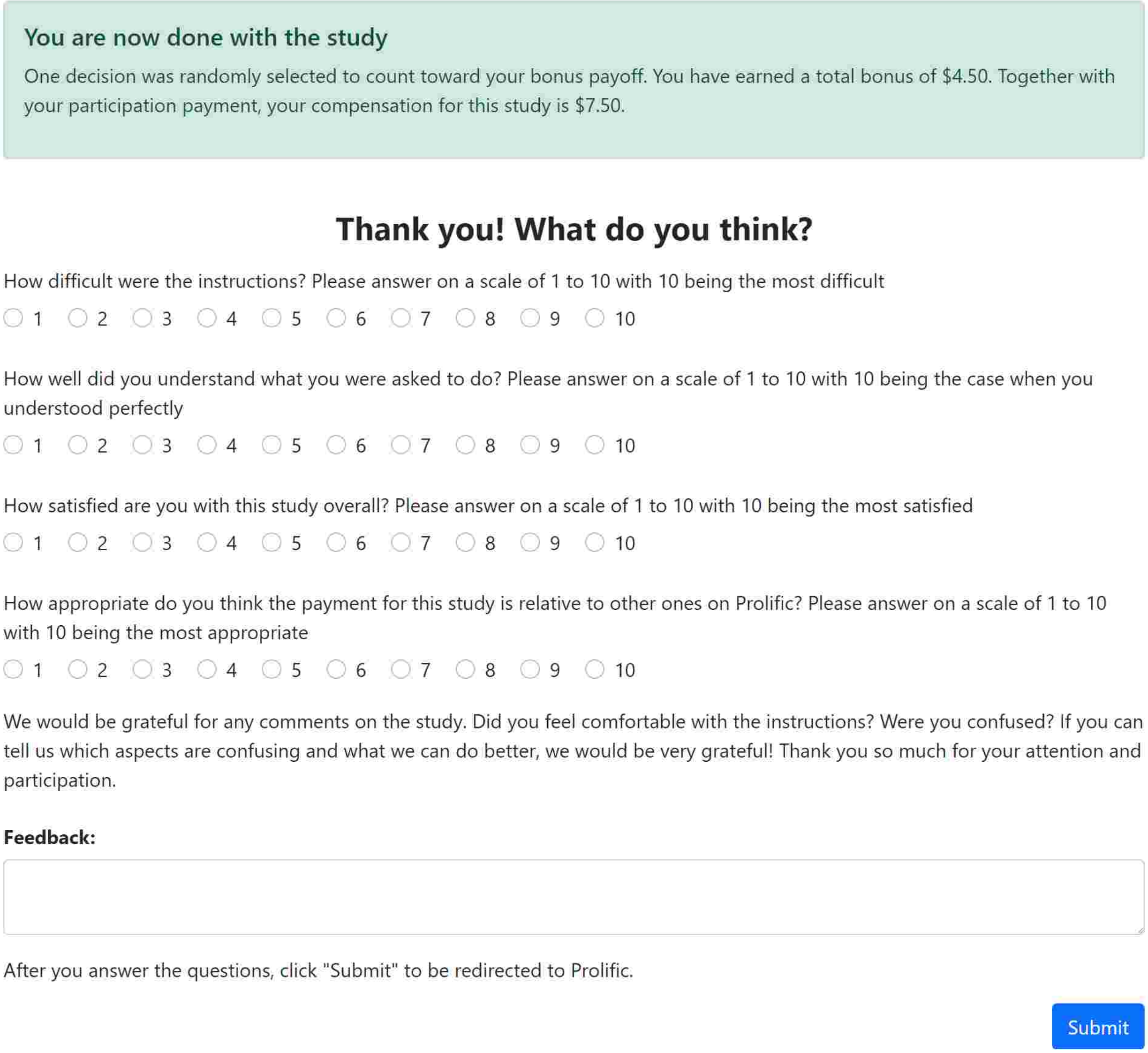
What is the highest degree you have received?

○ Less than high school degree
○ High school or equivalent including GED
○ Some college but no degree
○ Associate or technical degree in college (2-year)
○ Bachelor's degree in college (4-year)
○ Master's degree
○ Doctoral degree
○ Professional degree (JD, MD)

After you answer the questions, click "Next" to go to the last page of this study.

Figure A.35

Note: All treatments.



You are now done with the study

One decision was randomly selected to count toward your bonus payoff. You have earned a total bonus of $4.50. Together with your participation payment, your compensation for this study is $7.50.

Thank you! What do you think?

How difficult were the instructions? Please answer on a scale of 1 to 10 with 10 being the most difficult

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

How well did you understand what you were asked to do? Please answer on a scale of 1 to 10 with 10 being the case when you understood perfectly

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

How satisfied are you with this study overall? Please answer on a scale of 1 to 10 with 10 being the most satisfied

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

How appropriate do you think the payment for this study is relative to other ones on Prolific? Please answer on a scale of 1 to 10 with 10 being the most appropriate

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

We would be grateful for any comments on the study. Did you feel comfortable with the instructions? Were you confused? If you can tell us which aspects are confusing and what we can do better, we would be very grateful! Thank you so much for your attention and participation.

Feedback:

After you answer the questions, click "Submit" to be redirected to Prolific.

Figure A.36

Note: All treatments.