To Screen or Not to Screen: The Inference Cost of Policies*

Gonzalo Arrieta† Maxim Bakhtin‡

May 4, 2024

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. To investigate this, 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 the failures to anticipate inference and to plan the full strategy.

*We thank B. Douglas Bernheim, Lukas Bolte, John Conlon, Muriel Niederle, Kirby Nielsen, Alvin Roth, Jason Somerville, and seminar participants at ESA, BABEEW, BEESCuitS, and Stanford University for their helpful comments.

†garrieta@stanford.edu; Department of Economics, Stanford University

‡mbakhtin@stanford.edu; Department of Economics, Stanford University

I. Introduction

Screening is important and prevalent in many economic environments. In situations such as when looking for 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 about them. 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, the trade-off in screening is also essential for policy choices. A stringent policy may force everyone to behave the same way. This uniform behavior would leave no room for making inferences about people by observing their behavior. For example, if prisons strictly enforce their 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.

In this paper, we investigate whether people screen optimally and, if they do not, what causes their mistakes. The screening trade-off is complicated for several reasons. 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 various reasons. 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 uncertainty

inherent in assigning a difficult task to a new employee. Third, strategic considerations, such as uncertainty about the employee's response, may also prevent the employer from assigning a difficult task. Identifying whether people screen optimally has proven difficult because of these factors.

We hypothesize and experimentally test that people screen too little, even without these confounding factors. Our hypothesis stems from the intuition that the screening costs are usually immediate and evident when making a screening decision. At the same time, the benefits of the extra information are only realized in subsequent decisions and are harder to recognize. People 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 additional strong assumptions. The ideal setting needs to involve 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 not screening a mistake, revealing information at the initial stage must be optimal. To achieve this, we design a single-agent, single-period decision problem without uncertainty that satisfies these criteria. This 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 participants make two choices: one in part 1 and another in part 2. In part 1, they 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, 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. In 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 choose the Screening trial task know which computer is Good, but those who choose the Pooling trial task do not. This design makes information about the quality of computers valuable. The trade-off between a lower payoff from the Screening trial 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, and we focus on six of them, the parameters of which make the Screening trial task optimal.[1] We establish the Screening trial task's optimality as follows. We calibrate the payoffs to make the hiring choice stakes much larger than the trial task stakes, which is required to make screening worth it. The Screening task ensures a guaranteed high payoff, while the Pooling task induces a lottery with a much smaller expected payoff. Thus, participants should prefer the Screening task unless they are extremely risk-loving.[2] Furthermore, our design ensures no scope for time preferences or strategic reasoning to rationalize the Pooling task choice.

We use two treatments to identify suboptimal screening. In the *Baseline* treatment, participants solve the problem in the order described above. This treatment gives us our estimate of the share of mistakes. However, some mistakes in the *Baseline* treatment may come from the experimental noise. To estimate the amount of noise, we run a *Strategy Method* control treatment. This treatment helps participants as much as possible in making their decisions while maintaining the same experiment structure. The *Strategy Method* treatment allows us to estimate the share of mistakes that can be attributed to experimental noise, for example, driven by participants' inattentiveness or trembling hand errors. In this

[1] Including ten rounds in the experiment allows us to examine the persistence of our results. We mix six Screening-optimal rounds with four Pooling-optimal rounds to prevent participants from learning to choose the Screening task mechanically.

[2] Appendix Table A.1 lists the complete set of parameters used in all rounds.

treatment, participants solve the problem backward—starting with part 2 hiring choices conditional on the two possible tasks, and then making part 1 task choice. 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. If our hypothesis is true, the rate of mistakes in the *Baseline* treatment should be significantly higher than in the *Strategy Method* treatment.

We run the experiment online with 982 Prolific participants and find strong support for our hypothesis that people screen insufficiently. 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%. In subsequent rounds, after getting feedback and experience, *Baseline* participants learn the optimal strategy 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 are statistically significant. This result shows that insufficient screening is prevalent and does not fully disappear even with experience.

We investigate two driving mechanisms of insufficient screening. First, we propose that when choosing the task in our experiment, people fail to anticipate that they will be able to infer information about quality. We test the Failure to Anticipate Inference using the *Automatic Inference* treatment. This treatment is identical to the *Baseline* treatment, but participants automatically receive direct information about which computer is Good and which is Bad when choosing the Screening task. Participants know about this at the moment of choosing the trial task. This intervention eliminates the need to make inferences and thus serves as a diagnostic test for the Failure to Anticipate Inference mechanism.

Second, we test a different mechanism: whether people fail to plan their full strategy. The optimal choice of the trial task requires participants to plan their strategy for the sub-

sequent 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 trial task choice and the hiring choice— as a complete strategy. This treatment forces participants to think ahead and consider the entire strategy, thus identifying the Failure to Plan mechanism.

Results of the *Automatic Inference* and the *Plan* treatments show evidence for the proposed mechanisms. In the *Automatic Inference* treatment, the mistake rate is much lower than in *Baseline*—44% in round 1 and 22% on average in the subsequent rounds. Similarly, in the *Plan* treatment, the mistake rate is 36% in round 1 and 19% in the other rounds. Each mechanism accounts for about half of the mistakes in *Baseline* in the first round. In subsequent rounds, the mistake rates in both mechanism treatments are indistinguishable from the *Strategy Method* treatment. This result suggests that the interaction of the two mechanisms prevents complete learning. Even a partial intervention that eliminates one of the mechanisms is sufficient to help people 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. 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 the results remain the same on a subsample of participants who make zero errors. Third, the mistakes are not driven by suboptimal information use, which could justify choosing the Pooling task. Hiring mistakes are rare, 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. Lastly, the results are also robust to controlling for demographic characteristics and education.

We are closely related to the experimental literature on learning and bandit problems. This literature studies the exploration-exploitation trade-off—the choice between receiving a known payoff and an unknown but potentially higher payoff. A similar choice is present in our experiment. 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 (see Slivkins et al. 2019 for an 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 more naturalistic framework. In our study, participants tend to choose the Pooling task—they exploit even though exploration is optimal. Our second contribution to this literature is identifying the mechanisms behind under-exploration — Failure to Anticipate Inference and Failure to Plan. Merlo and Schotter 1999 and Merlo and Schotter 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 and contingent reasoning. People fail to make the right choice in strategic settings, where it requires thinking about others' actions and beliefs and learning from others' behavior (Milgrom and Roberts 1986; Eyster and Rabin 2005; Esponda and Pouzo 2017; Dal Bó et al. 2018; Eyster 2019). We contribute by showing that some of these results are not necessarily tied to a strategic setting. 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. This mechanism shows

that the difficulty of backward induction extends beyond strategic settings. Literature on failures of contingent reasoning shows that people make mistakes when they need to evaluate multiple hypothetical scenarios (Esponda and Vespa 2014; Martínez-Marquina et al. 2019). Our paper suggests that the difficulty of contingent reasoning applies not only to exogenous hypothetical events but also to hypothetical events arising from an individual's own choices (Niederle and Vespa, 2023).

The implications of failures to anticipate inference and to plan also relate to other strands of literature. We speak to the literature on the unintended consequences of policies (Bitler and Karoly 2015; Nandi and Laxminarayan 2016). It has been shown 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 designers 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. Stereotypes and wrong beliefs have been shown to be persistent (Snyder 1981; Skrypnek and Snyder 1982; Babcock et al. 2017). Our paper suggests one mechanism for this persistence. A society may create strong incentives for uniform behavior. As a result, the stereotyped group cannot disprove the stereotype, leading to its persistence. Moreover, Bernheim and Whinston 1998 show that incomplete contracts may be optimal if they allow learning more about the counterparty than complete contracts. Our results suggest that people are likely to ignore this channel and write too restrictive contracts. The idea of strong incentives impeding learning is also present in the literature on career concerns (Scharfstein and Stein 1990; Chevalier and Ellison 1999).

The rest of the paper is structured as follows. Section II describes the experimental design. Section III summarizes the main results. Section IV tests for the mechanisms. Section V shows the robustness of the results. Section VI discusses some implications of the results and concludes.

II. Experimental Design

We design an experiment to test whether people screen too little. 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 to make 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. The participants' policy choice reveals whether they suboptimally choose the policy that does not provide information—i.e., do they 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, the participants' social preferences toward human candidates could affect their choice of trial task and hiring decision. The direction of this confounder is ambiguous. Third, we need to ensure consistent performance of candidates across tasks—this is what defines the Good and the Bad candidates. Human candidates' performance is likely too inconsistent. As a result, the trial task would 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.E and IV.A. 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 stages: the trial task stage, which is the main decision of interest, and the hiring stage. Figure I illustrates the structure of each round. The first stage, the trial task stage, allows the participant to infer which candidate is better. The second stage, the hiring stage, makes this information valuable. Specifically, two computers, Good and Bad, generate a payoff for the participant by solving tasks. There are two tasks: Pooling and Screening. In the Pooling task, both computers generate the same high payoff. In the Screening task, the Good computer generates a high payoff, while the Bad computer generates a low payoff. Thus, the participant can infer the quality of each computer by observing the payoffs they generate in the Screening task. This inference exercise is crucial in our design.

Choosing the Screening task provides information, but it comes at a cost. In the Pooling task, both computers generate a high payoff. In the Screening task, only the Good computer generates a high payoff, while the Bad one generates a low payoff. Thus, the participant faces a trade-off between receiving a lower payoff from the Screening trial task and learning the computers' quality.

After choosing the Pooling or the Screening trial task, the participant enters the hiring

stage, in which knowing the computers' quality is valuable. At the hiring stage, the task the computers solve is always the Screening task. The participant then chooses which of the two computers to hire for another Screening task. In this decision, the stakes are much higher than in the trial task. The higher stakes justify forgoing a part of the payoff in the trial task to learn which computer is Good. We discuss the optimal choice of the trial task in subsection II.C.

Participants face ten rounds of this problem with varying payoff parameters. The first and last rounds are the same for all participants, and rounds 2–9 appear in random order. Six parameterizations, including the first and last rounds, follow the description above. Under these parameters, it is optimal to choose the Screening trial task. For example, we use the following values in the first round. In the Pooling trial task, both computers generate \$0.05. In the Screening trial task, the Good computer generates \$0.05, while the Bad computer generates \$0.00. In the second stage, the Good computer generates \$4.30, while the Bad computer generates only \$0.05. The other four parameterizations make it optimal to choose the Pooling trial task. These parameterizations appear in random order in rounds 2–9. While our primary interest is the Screening-optimal rounds, we include Pooling-optimal rounds to prevent participants from mechanically choosing the Screening trial task without regard for the payoffs. Appendix Table A.1 summarizes the parameter values in the ten parameterizations. 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.

At the beginning of the experiment, we 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 hiring stage. We employ rigorous comprehension checks to verify that the partic-

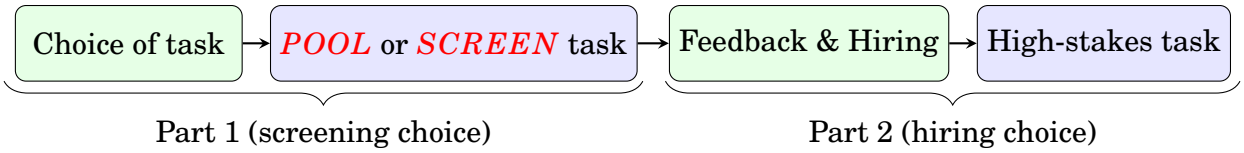


Figure I: Diagram of experiment

ipants understand the setup.[3]

II.C. Pooling is a mistake

The only factor to consider in determining whether Pooling is a mistake is the participant's risk preferences because the two trial tasks induce two different lotteries. We show that participants' risk preferences are such that choosing the Pooling trial task in the Screening-optimal rounds is 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 also do not matter for the optimal choice.

All six Screening-optimal rounds have similar payoff structures. We use the first round, which is the same for everyone, to illustrate the payoff consequences. If the participant chooses the Screening trial task and hires the Good computer, she gets \$4.35 with certainty. If the participant chooses the Pooling trial task, she faces a lottery with equally likely payoffs of \$4.40 and \$0.15. The Screening trial task is optimal unless the participant is sufficiently risk-loving.[4] In this and other Screening-optimal rounds, the Pooling task reduces the expected payoff by at least \$2 compared to the Screening task.

We elicit participants' risk preferences directly to diagnose the strength of our assumption that the Screening trial task is optimal. Assuming that a direct choice over lotteries better captures participants' risk preferences, this exercise supports our claim that choosing the Pooling task is a mistake.[5] After the main part of the experiment, we ask the participants

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

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

[5]While we find this to be a reasonable assumption, and approaches like this one are not rare in experimen-

to choose between two lotteries for each of the six Screening-optimal parameterizations: one is a fixed payoff resulting from the Screening task and hiring the Good computer, and the other is a lottery resulting from the Pooling task and hiring a computer randomly. Thus, we elicit participants' preferences over the induced lotteries in each parameterization. Figure A.1 in the appendix summarizes the results. Across all six parameterizations, less than 3% choose the lottery that corresponds to the Pooling trial task.

II.D. 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 on Prolific 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*.[6] The average payoff is \$6.79, 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. Figure A.2 in the appendix shows the distribution of mistakes on the understanding questions. Restricting the sample to participants who make zero mistakes does not affect the

tal economics, Bernheim and Taubinsky 2018 discusses how it can suffer from the circularity trap, by which "we identify bias by looking for choices that conflict with true preferences while inferring true preferences from unbiased choices."

[6]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.2 in the appendix summarizes demographic characteristics by treatment.

results. 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 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 V).

II.E. Baseline and Strategy Method Treatments

We start with two treatments that show that participants do not screen optimally. The *Strategy Method* treatment serves as a control and establishes a benchmark for the amount of noise in the participants' answers.[7] The *Baseline* treatment shows that participants consistently make wrong choices.

The *Baseline* treatment identifies insufficient screening in a natural environment. We design it 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. We interpret these errors as experimental noise, which 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 need to separate the two to measure the true prevalence of screening mistakes. We modify the *Baseline* treatment to measure a benchmark noise level.

To measure this noise benchmark, we design a control treatment called the *Strategy Method* treatment. This control treatment is meant to help participants as much as possible while keeping the structure of the experiment the same. We use the share of participants

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

who still make suboptimal choices in such a scenario as a measure of noise. In the *Strategy Method* treatment, participants submit their whole strategy, which requires solving the problem backward. They start with the hiring decision and choose which computer to hire in each scenario—i.e., when they know the computers' quality and when they do not. At this stage, the participants see the labels for the Good and Bad computers following the Screening task. Next, they complete the first stage, choosing the 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.

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.

Choosing the Pooling trial 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. We expect participants to learn over time from the feedback and experience. Therefore, we present results separately for round 1 and rounds 2–10. In rounds 2–10, participants' choices are affected by learning.

In the first round, most participants make the mistake of forgoing valuable information. 68% of participants choose the Pooling trial task in the *Baseline* treatment. In the *Strategy Method* treatment, only 18% choose the Pooling task. The difference is statistically significant ($p < 0.001$).

There is substantial learning from experience, but mistakes disappear only partially. In rounds 2–10, the rate of mistakes reduces to 34% in *Baseline*. It is still significantly higher

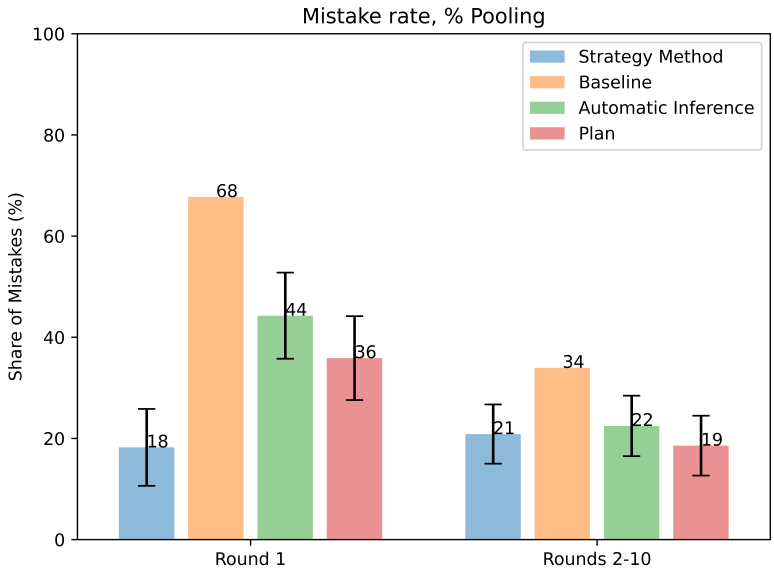


Figure II: Mistake rates in all treatments in round 1 and rounds 2–10. 95% confidence intervals are shown for the difference from the *Baseline* treatment and calculated using standard errors clustered at the participant level. Sample sizes are 251 for *Baseline*, 244 for *Strategy Method*, 244 for *Automatic Inference*, and 251 for *Plan* treatments.

than the rate of mistakes in the Strategy treatment, which is 21%. Figure II summarizes these results. Most of the learning happens after round 1. Figure A.3 shows the rate of mistakes by round. It stays flat across all rounds in the *Strategy Method* treatment and 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 without accounting for multiple hypothesis testing.

These mistakes are substantial in terms of payoff consequences. By not screening, participants leave on the table at least $2, which represents 53% of their average bonus payment.

IV. Mechanisms: Failure to Anticipate Inference and Failure to Plan

In this section, we investigate the mechanisms behind insufficient screening. By our experimental design, the mistake of choosing the Pooling trial task cannot be attributed to standard explanations of time inconsistency or risk aversion. It also cannot be caused by strategic reasoning mistakes or the difficulty of thinking through multiple states of the

world. To solve the problem correctly, participants must recognize the informational content in the computers' performance on the Screening task and use it to plan the strategy. We separately test for both — whether they fail to anticipate the inference they will be able to make and whether they fail to plan their strategy. We call these mechanisms Failure to Anticipate Inference and 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 we make participants choose their full strategy from the beginning. These interventions substantially reduce the mistakes, which supports Failure to Anticipate Inference and Failure to Plan as mechanisms.

IV.A. Automatic Inference and Plan Treatments

The first hypothesized mechanism is the Failure to Anticipate Inference. This mechanism suggests that participants do not know they will be able to make inferences in the future: they do not realize that the Screening trial 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. We introduce the *Automatic Inference* treatment, a light-touch intervention to the *Baseline* treatment, to test for Failure to Anticipate Inference.

The *Automatic Inference* 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 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 at the trial task stage when choosing between the Screening and Pooling tasks. 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.

The second hypothesized mechanism is Failure to Plan. Under this mechanism, 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.

The *Plan* 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) Picking the Screening trial task and hiring the computer that produces the larger amount, (ii) picking the Screening trial task and hiring the computer that produces the smaller amount, and (iii) picking the Pooling trial task and hiring one the computers chosen randomly. This elicitation forces participants to consider the whole problem and plan their entire strategy. If the Failure to Plan mechanism is correct, this planning tool should also reduce the rate of mistakes relative to the *Baseline* treatment.

IV.B. Results: Helping with Inference and Planning Reduces Mistakes

The *Automatic Inference* and *Plan* treatments substantially reduce the mistake rate. This result suggests that the Failure to Anticipate Inference and Failure to Plan mechanisms are important drivers of insufficient screening. Participants do not understand how to extract useful information in the future and do not plan the entire strategy.

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* and the *Strategy Method* treatments. Furthermore, we observe complete learning to the same level as in the *Strategy Method* treatment. Appendix Figure A.4 shows that the rate of mistakes stays about the same starting with round 2. The difference in the mistake rate from the *Baseline* is highly significant

($p < 0.001$) in round 1 and in rounds 2–10 combined. 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. In the first round, 36% of the participants choose the Pooling trial task, which is similar to the *Automatic Inference* treatment[8]. In the subsequent rounds, participants learn completely — the mistake rate in rounds 2–10 is similar to the *Strategy Method* treatment. Appendix Figure A.5 shows that the rate of mistakes stays about the same starting with round 2. The difference in the mistake rate from the *Baseline* is significant ($p < 0.001$) in round 1 and in rounds 2–10 combined. Furthermore, comparing round by round, the mistake rate is significantly lower in the *Plan* treatment than in the *Baseline* in each round separately. Thus, the *Plan* treatment suggests that the hypothesized Failure to Plan mechanism is important too. Participants do not seem to plan their entire strategy.

The difference in the average number of mistakes between treatments is driven largely by the different shapes of the mistake distributions. For each participant, we calculate the number of mistakes they make in the Screening-optimal rounds (i.e., choosing the Pooling trial task). Aggregating over all participants within a given treatment, we plot the histograms of the total number of mistakes in Appendix Figure A.6. 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.

The group of *Baseline* participants who always choose the Pooling task mainly drives the persistence of insufficient screening. These participants never learn that screening is optimal and consistently avoid screening throughout the experiment. If we remove from

[8]The difference between *Automatic Inference* and *Plan* treatments is insignificant at 5% level.

the sample participants who never learn, 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 I). This result shows that some participants struggle to learn the optimal screening choice without aid. As a result, the mistake rate is more persistent in the *Baseline* treatment than in the other treatments that eliminate one or several mechanisms.

The results of the two mechanism treatments suggest that participants fail to anticipate inference and plan their strategies. The two mechanisms seem to be similarly important drivers of insufficient screening. Therefore, both interventions help participants screen optimally.

V. Robustness of Results

We conduct multiple tests to verify the robustness of our results. All results are summarized in Table I.

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. Excluding participants who prefer at least one lottery induced by the Pooling task does not affect our results (Table I, column (2)).

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

We record the number of errors on the understanding questions and plot its distribution in Appendix Figure A.2. Most participants make few errors. In particular, 89% of participants make at most one error, and 56% make zero errors. The small number of understanding

errors highlights that confusion or misunderstanding was not an issue for most participants. The average number of mistakes 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.7). Therefore, misunderstanding is unlikely to drive the treatment effects.

We further show that the results are robust to misunderstanding by focusing on the participants with perfect understanding—the 546 participants who make zero errors on the understanding questions. The results remain the same as in the full sample, as shown in Table I, column (3).

Participants in the *Baseline* treatment need to understand that they will observe how much each computer produces before they make the hiring choice. If they fail to understand this, the Screening trial 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. It could shut down the mechanisms that drive insufficient screening. To avoid interfering with the mechanisms, we first elicit participants' first trial task choice and only then ask the understanding question about whether they will observe how much each computer produces. As a result, the understanding question does not affect round 1, but it may bias the mistake rate in rounds 2–10 in *Baseline* down, which goes against our hypothesis. We find that 71% answer correctly—they know that they will observe how much each computer produces before the hiring choice. Restricting the sample to only those participants who answer correctly does not affect the results, as we show in column (4) of Table I.

Optimal use of information. Choosing the Screening trial task is optimal only if the participants use the information it reveals correctly. Therefore, to claim that choosing the Pooling task is a mistake, we need to show that participants who choose the Pooling task would still be able to use the information about quality if they got it. An ideal test for this would be observing participants' counterfactual hiring choices after the Screening task.

We approximate this counterfactual scenario by looking at participants' actual choices after choosing the Screening task. 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 the participants hire the Good computer after choosing the Screening task (Appendix Figure A.8). The few hiring mistakes are unlikely to drive the differences in the choice to screen.

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. We restrict the sample to those participants who have chosen the Screening task at least once and have always hired the Good computer after that. We interpret this sample as the participants who use the information correctly. Note that it automatically excludes everyone who always chooses the Pooling task. 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 the 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 I, column (6)). This result is expected because the never-learners drive the persistence of insufficient screening in the full *Baseline* sample and are excluded from this subsample.[9]

Individual characteristics. We rule out that differences in the samples drive the treatment effects. We elicit participants' self-reported age, gender, race, and education. The

[9]An alternative approach is to exogenously reverse the trial task choice from Pooling to Screening, to have participants who choose Pooling 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 exogenous reversal, which could drive mistakes. Second, it provides a counterfactual only probabilistically, in a small share of rounds.

Table I: 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

(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: ***

treatments are balanced on percent female, percent White, and percent college-educated, but the *Plan* participants are slightly younger than the *Baseline* participants (Appendix Table A.2). Column (5) of Table I shows the results are robust to controlling for age, gender, race, and education level.

VI. Conclusion

We experimentally show that people do not screen enough, even without time inconsistency, risk aversion, and strategic interactions. Suboptimal screening significantly diminishes, yet

persists, with experience and feedback. We further show that two mechanisms drive insufficient screening. First, participants fail to anticipate that they can infer the computers' quality. Second, they fail to plan their entire strategy. We run two additional treatments that shut down each of the mechanisms. Each of these interventions 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 policies. Policies that create strong incentives induce uniform behavior, which limits the amount of information that can be learned from observations. 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 was 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. Furthermore, it would be interesting to study cases where the party making the inference differs from the one making the screening choice to shed light on how this affects the trade-off. As a second direction, it would be interesting to empirically quantify the inference costs of policies in naturally occurring settings, including contexts where the policy choice is paternalistic. This direction would show the external relevance of our experimen-

tal results. An important third direction is studying the Failure to Plan in more detail. This mechanism can affect choices in any dynamic problem. Therefore, it is crucial to understand to what extent it could explain known mistakes in dynamic choices. A fourth, related, direction is investigating the effects and the demand for planning in dynamic environments. Usually, committing to a future strategy is seen as a way to overcome time inconsistency. In our experiment, however, there is no scope for time preferences. Nevertheless, participants value the ability to plan and benefit from it. This result calls for further investigating commitment devices and how they may help make optimal dynamic choices.

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.

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.

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.

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.

CHEVALIER, J. AND G. ELLISON (1999): "Career concerns of mutual fund managers," *The Quarterly Journal of Economics*, 114, 389–432.

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.

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.

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.

SCHARFSTEIN, D. S. AND J. C. STEIN (1990): "Herd behavior and investment," *The American economic review*, 465–479.

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.

A. Appendix

A.A. Tables and Figures

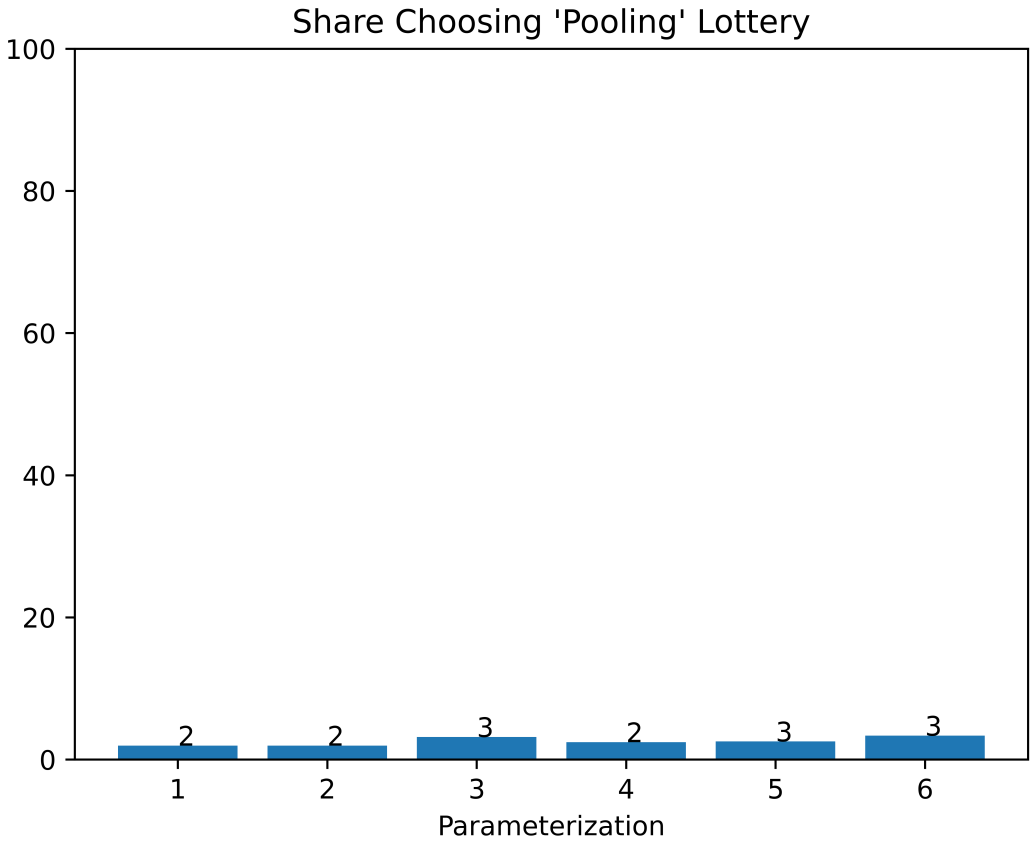


Figure A.1: The share of participants who choose the lottery induced by the Pooling trial task

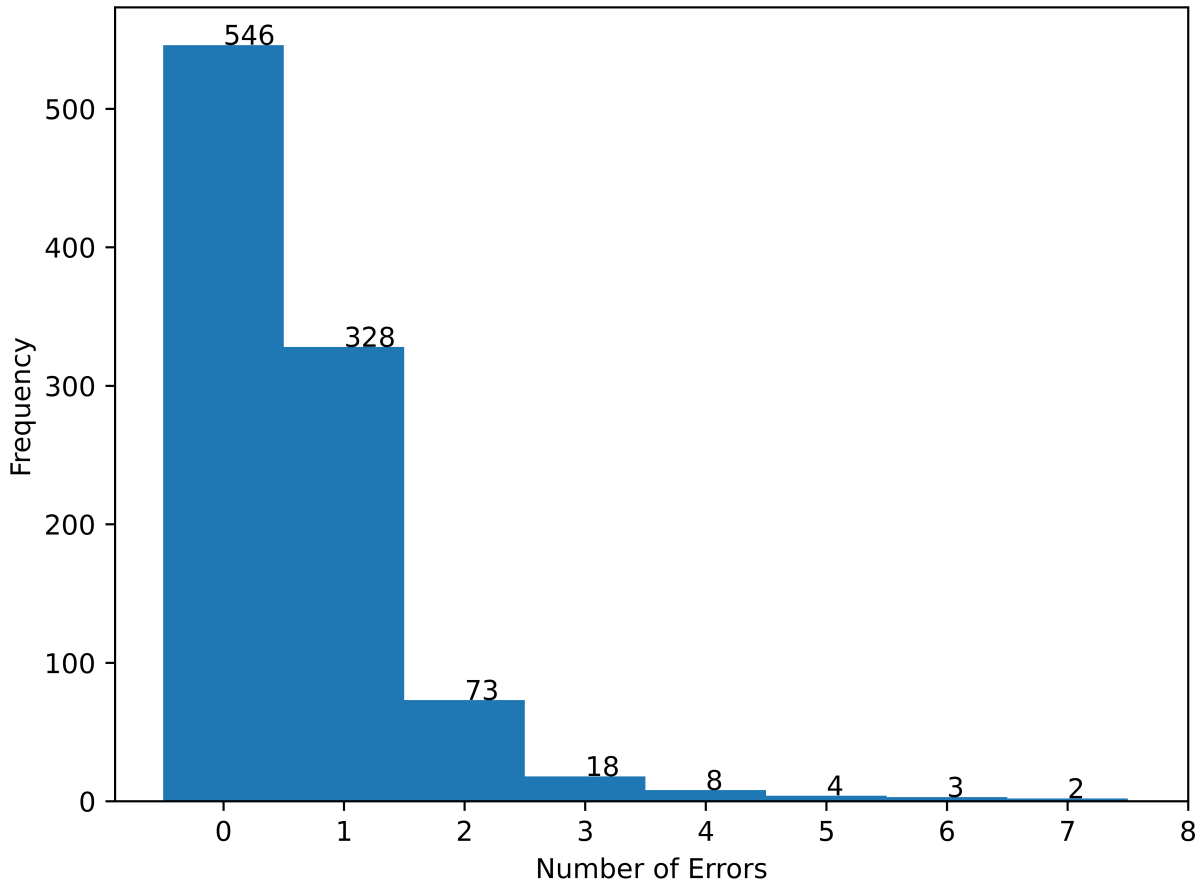


Figure A.2: The distribution of the number of errors on the understanding questions.

	Part 1			Part 2		
	POOL	*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 A.1: Summary of payoff parameters for the ten parameterizations. All values are in USD terms.

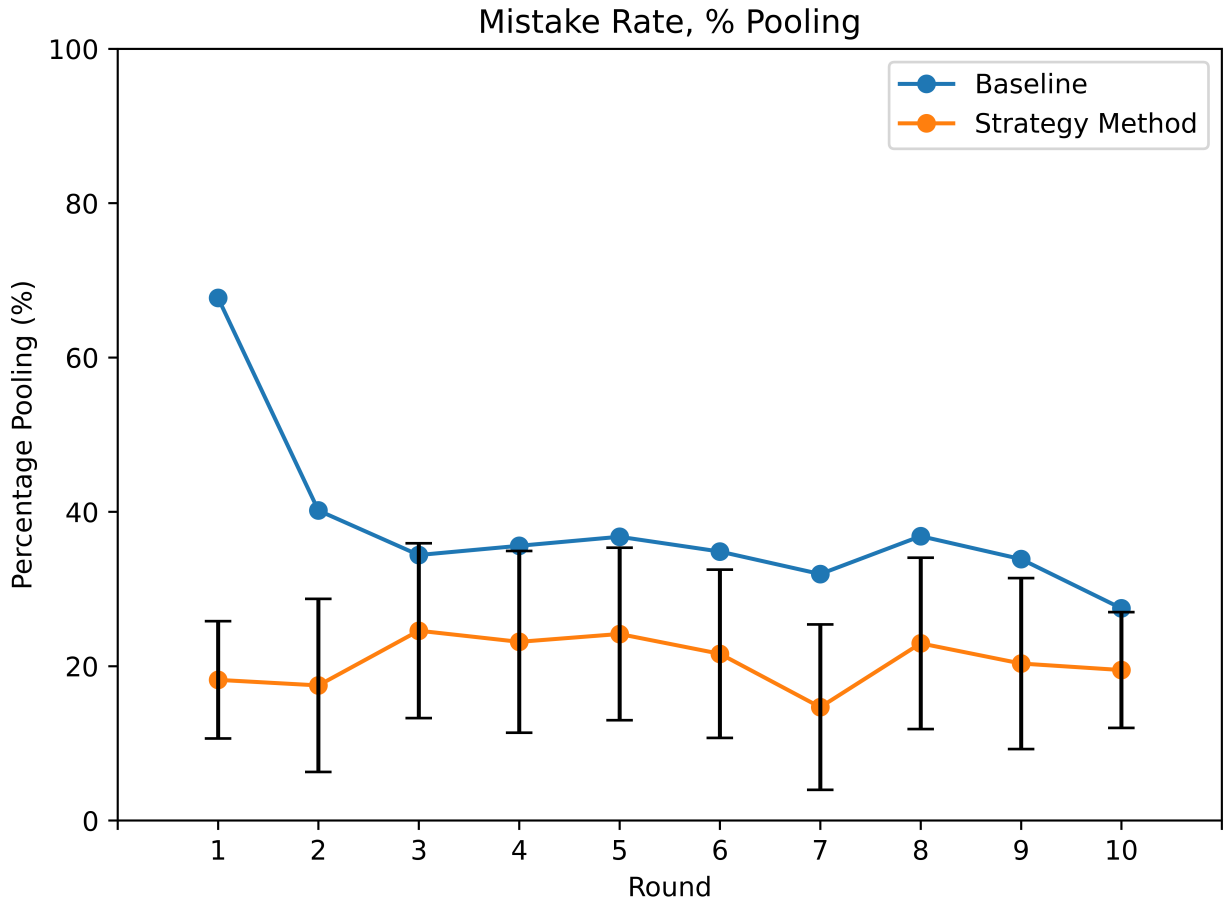


Figure A.3: Rate of mistakes across rounds with Screening-optimal parameterizations for *Baseline* and *Strategy Method* treatments.

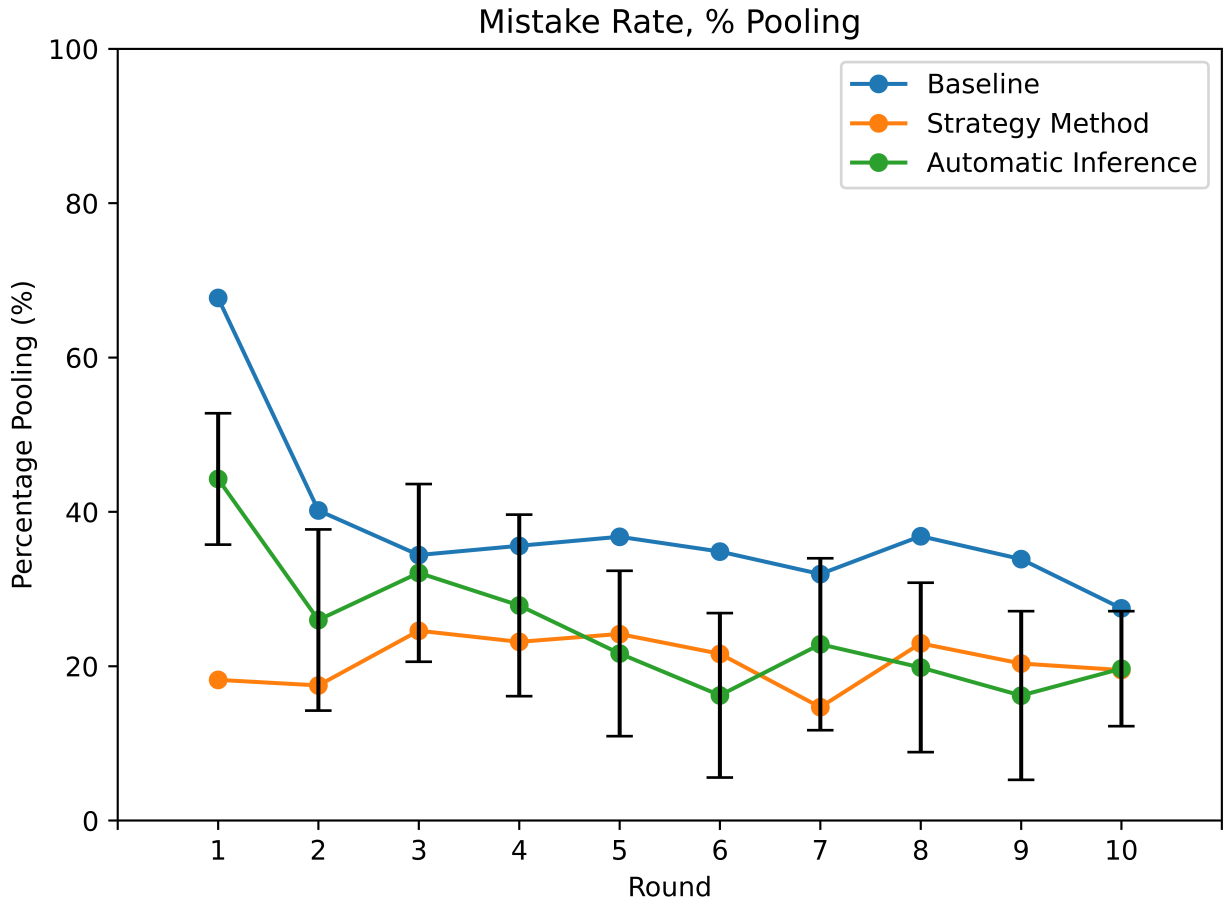


Figure A.4: Rate of mistakes across rounds with Screening-optimal parameterizations for *Baseline*, *Strategy Method*, and *Automatic Inference* treatments.

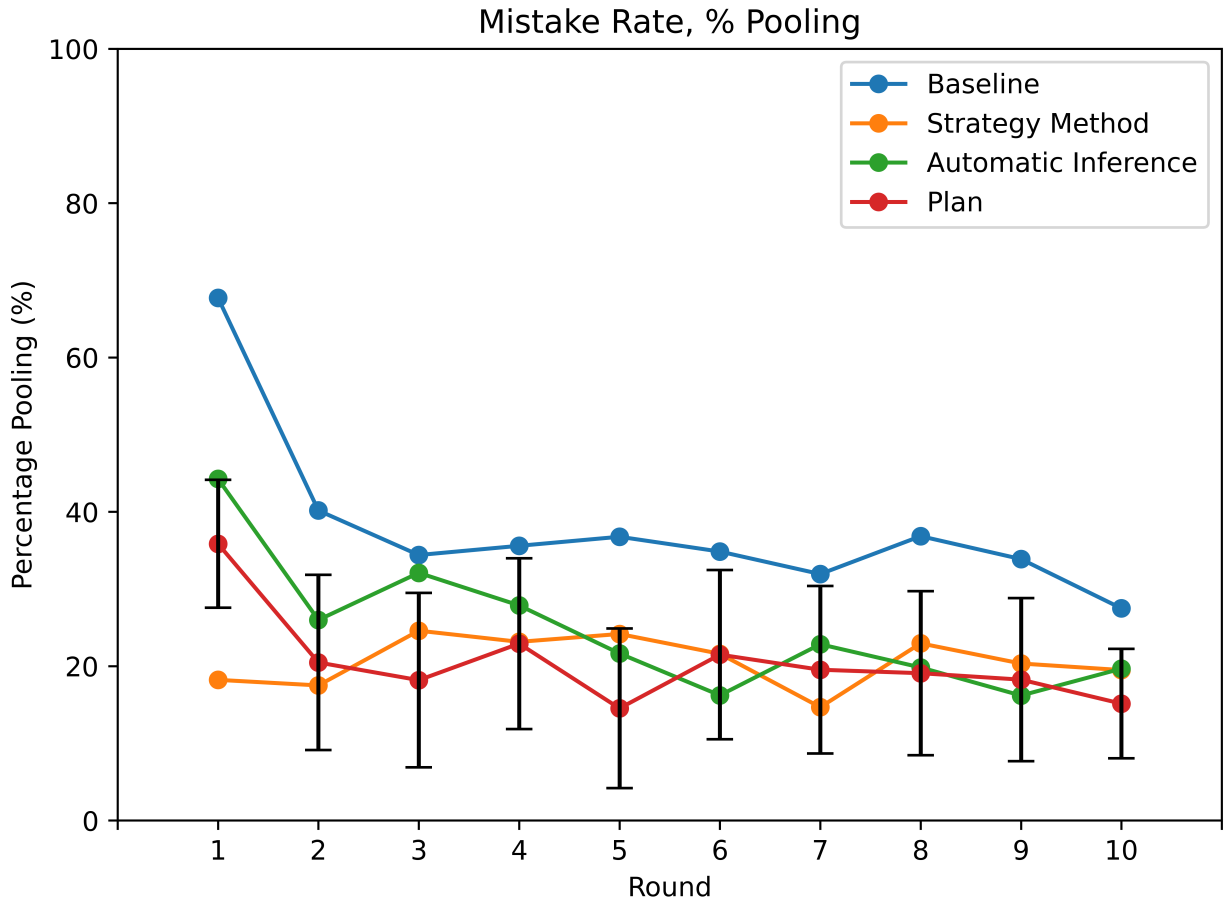


Figure A.5: Rate of mistakes across rounds with Screening-optimal parameterizations for all treatments.

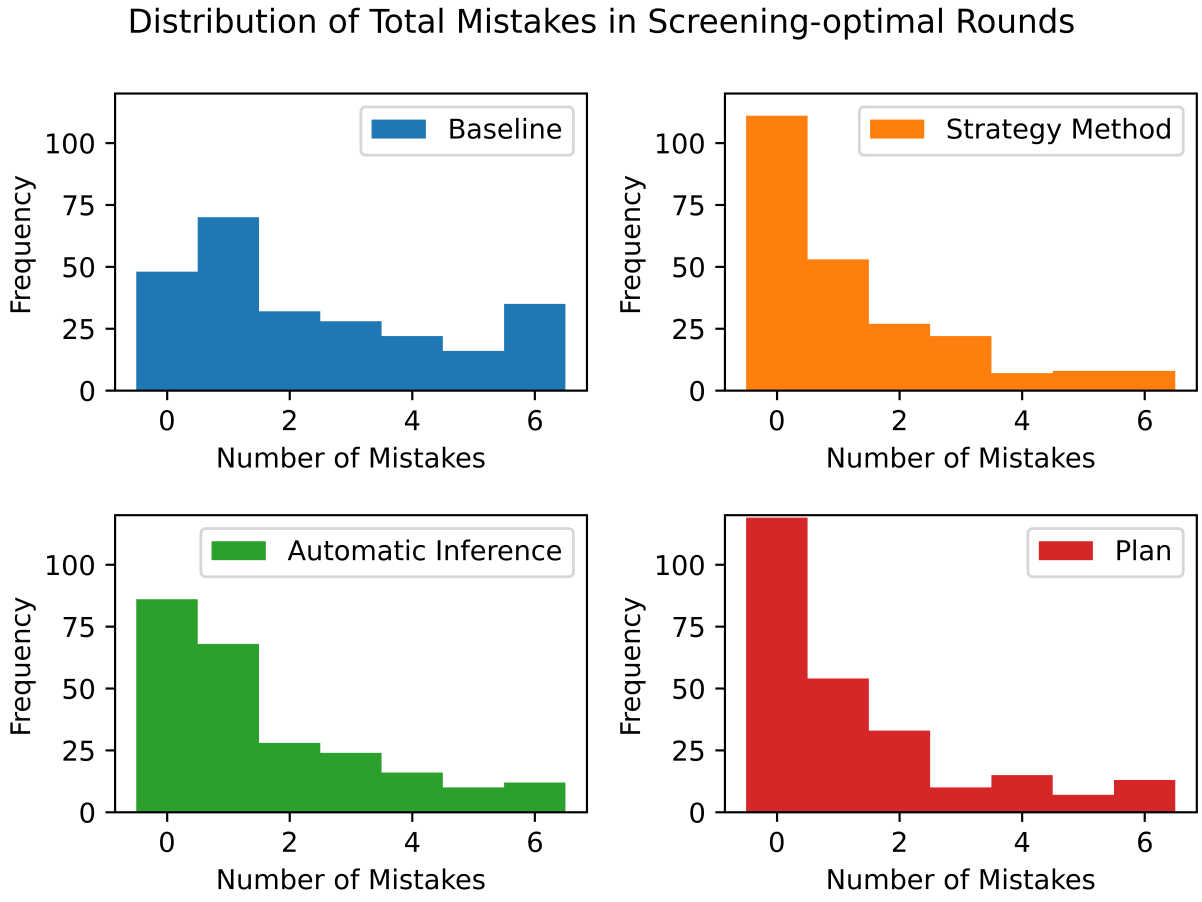


Figure A.6: The distribution of total screening mistakes participants make by treatment.

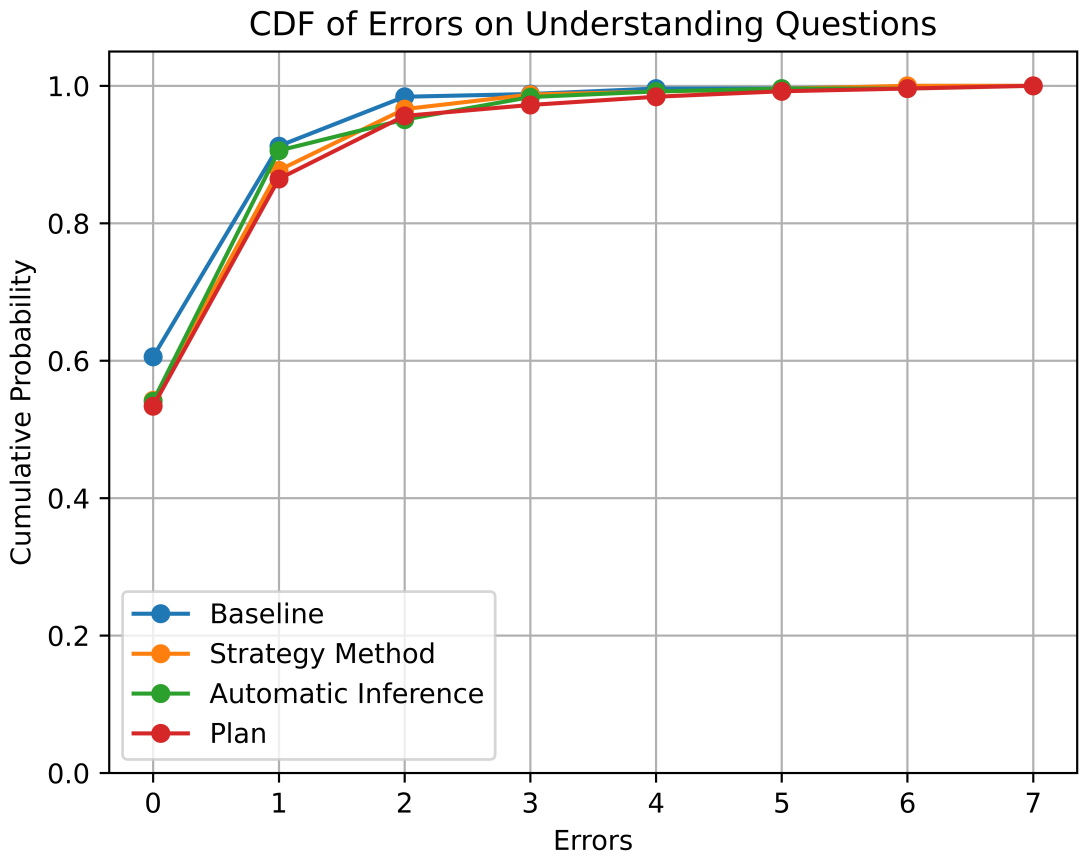


Figure A.7: The CDF of total errors on the understanding questions participants make by treatment.

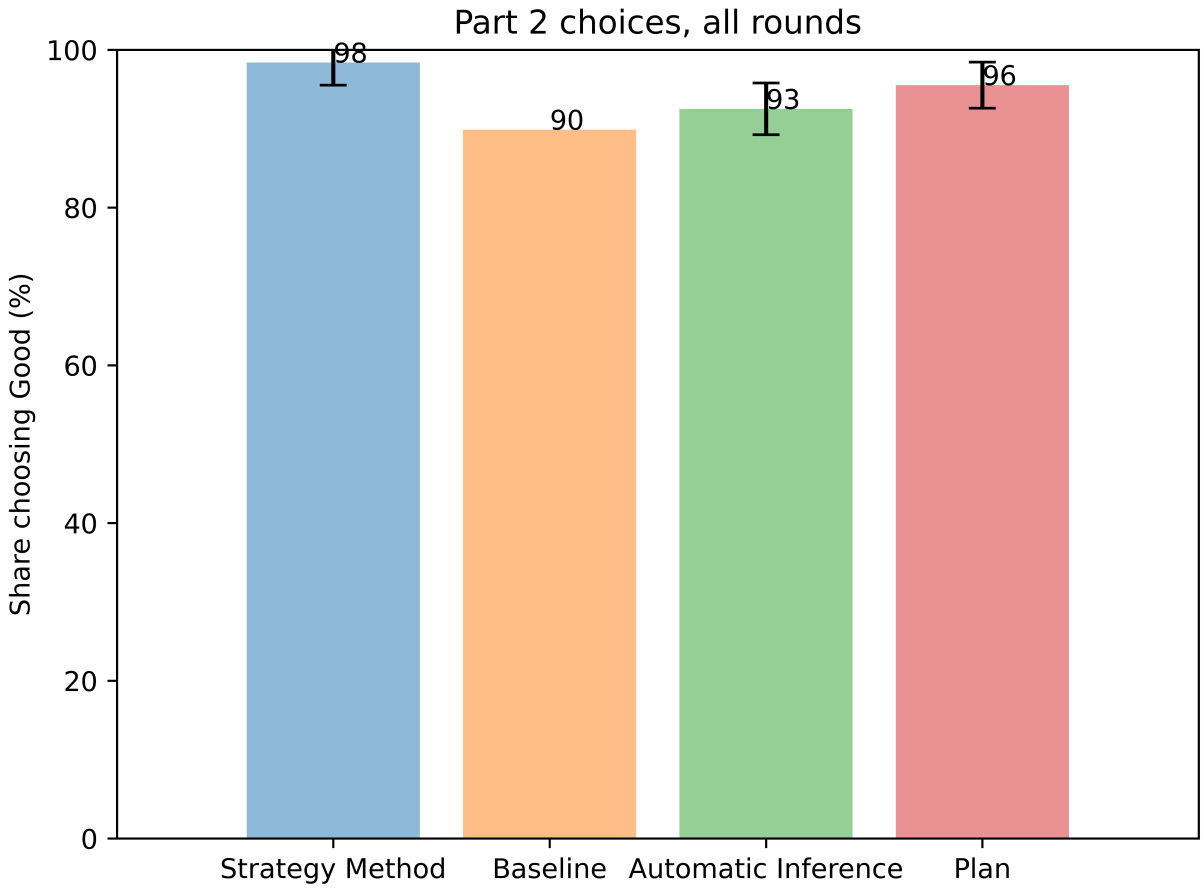


Figure A.8: Rate of hiring the Good computer in part 2 after choosing the Screening trial task by treatment.

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

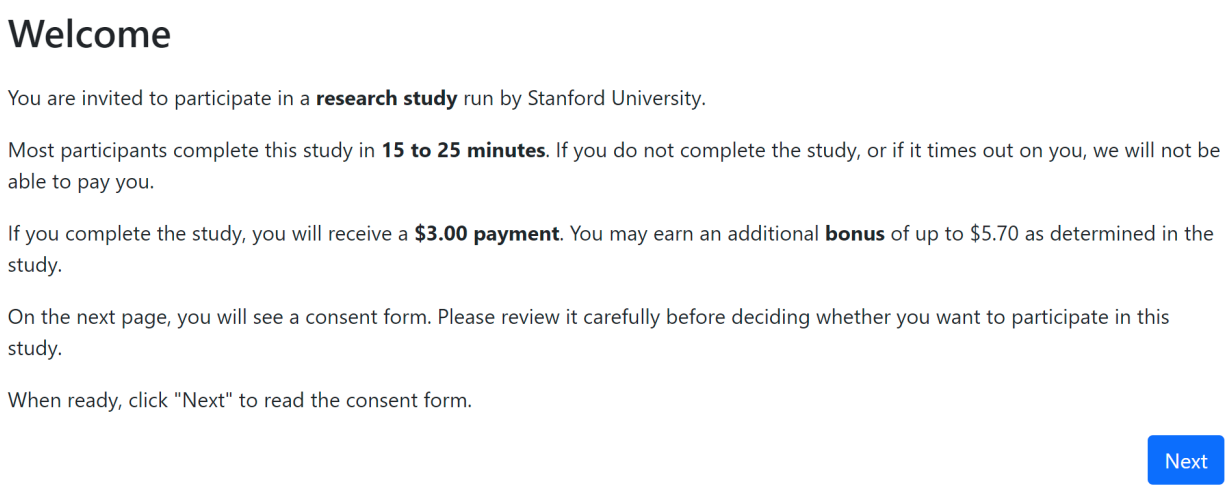


Figure A.9
Note: All treatments.

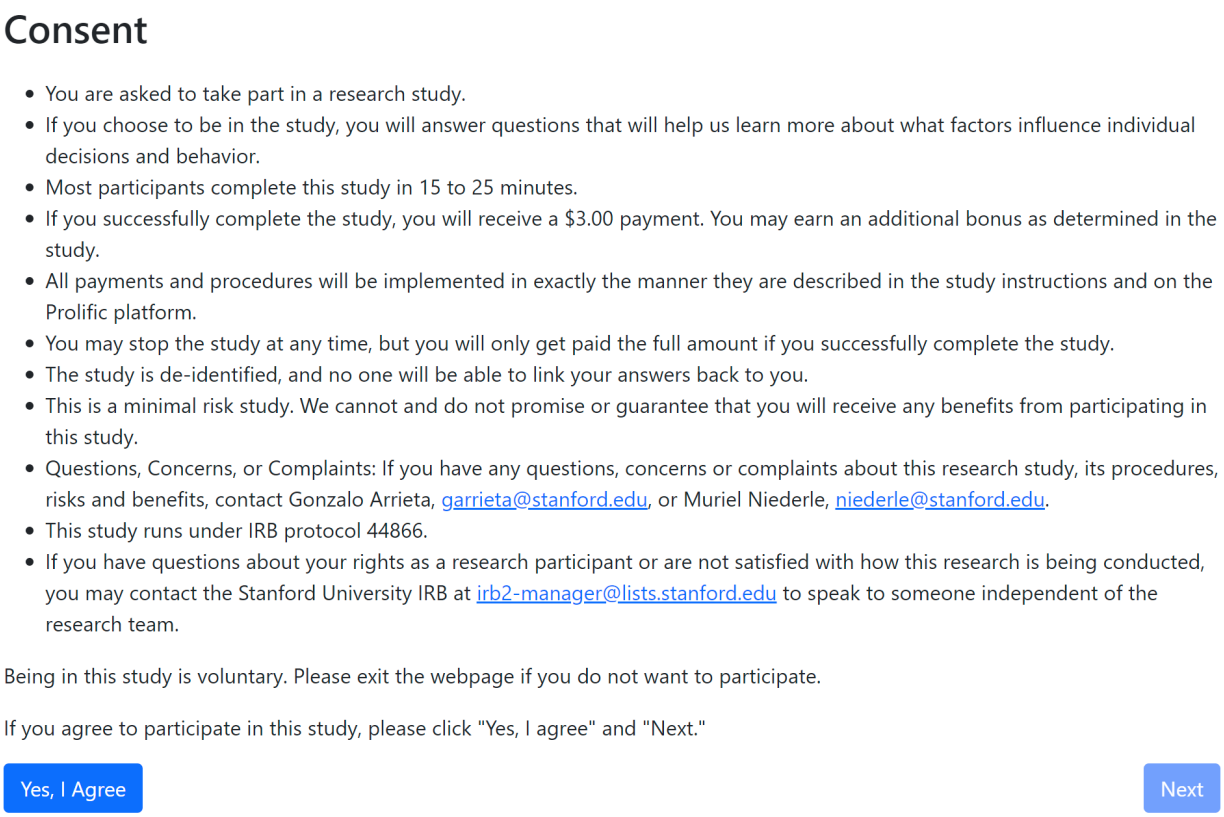


Figure A.10
Note: All treatments.

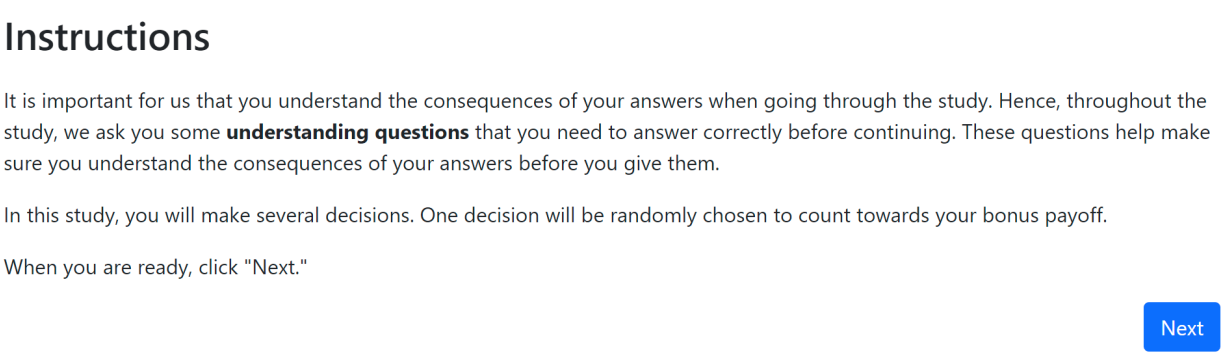


Figure A.11
Note: All treatments.

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.

Next

Figure A.12
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.

Next

Figure A.13
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.

Next

Figure A.14
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.

Next

Figure A.15
Note: *Plan* treatment.

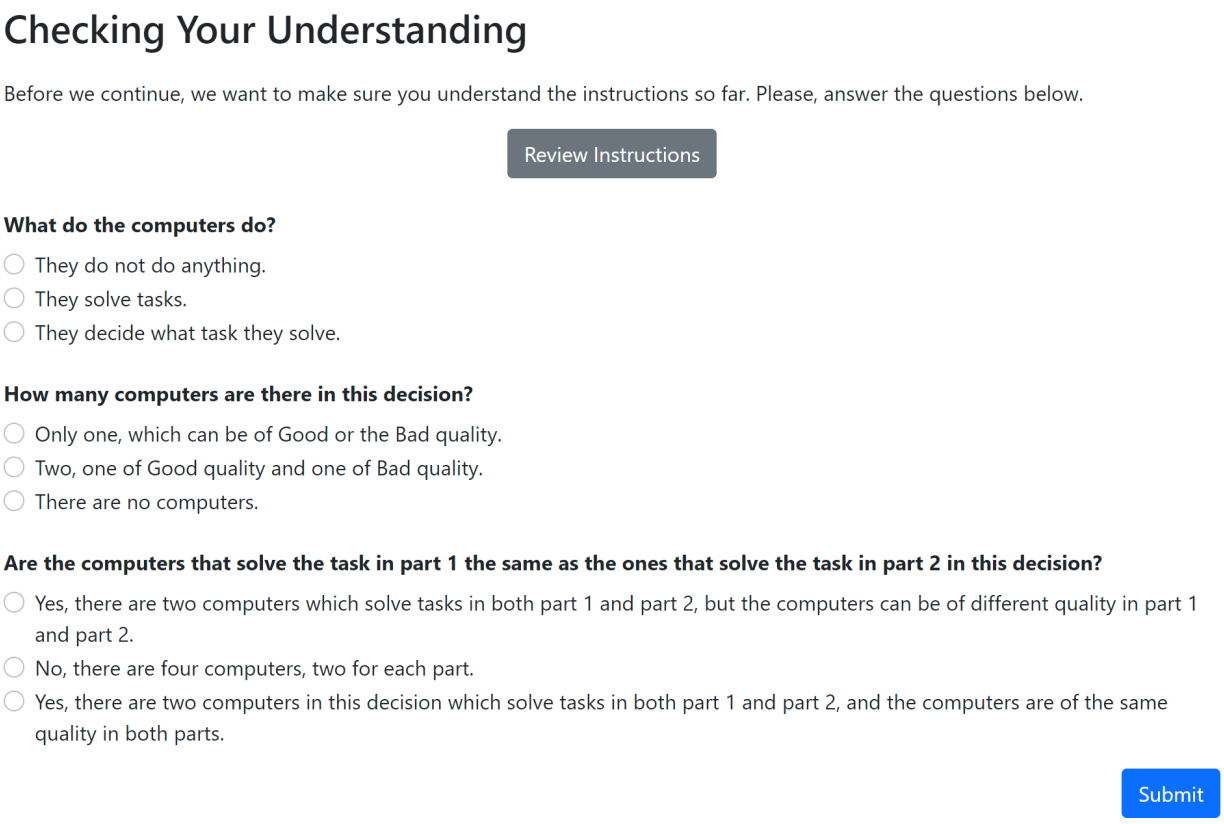


Figure A.16
Note: All treatments.

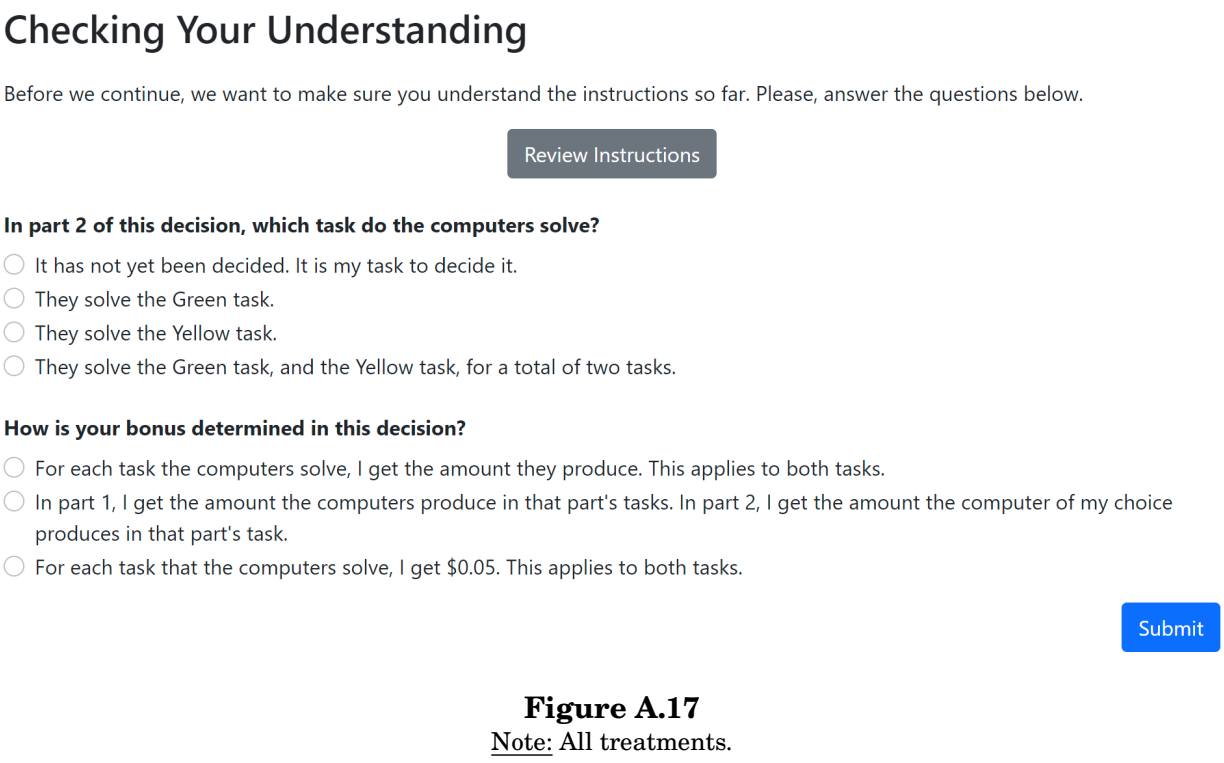


Figure A.17
Note: All treatments.

Main decision: Baseline treatment

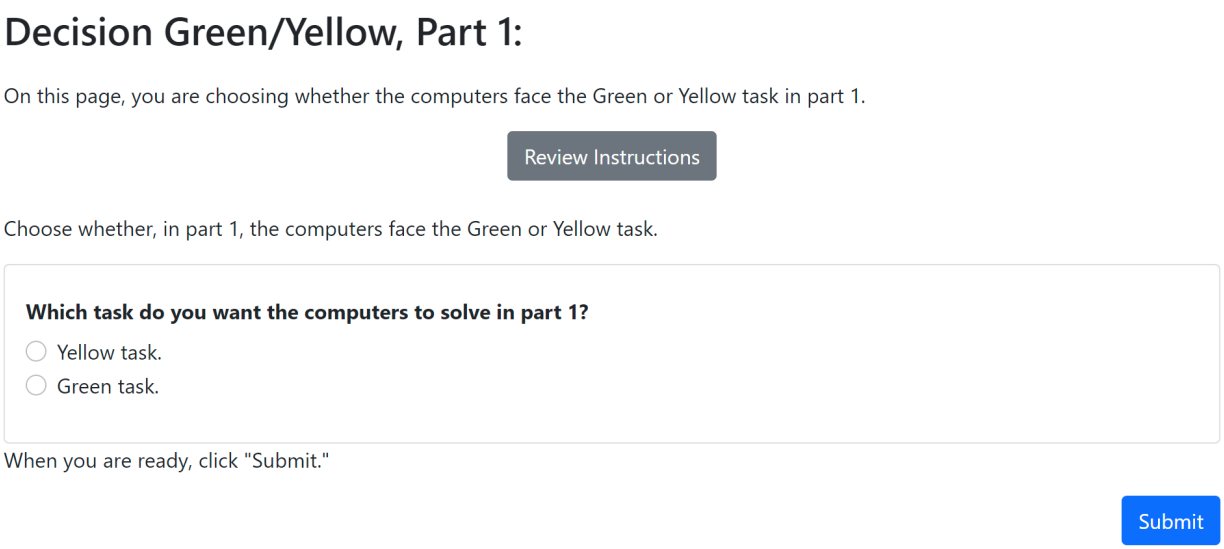


Figure A.18
Note: *Baseline* treatment.

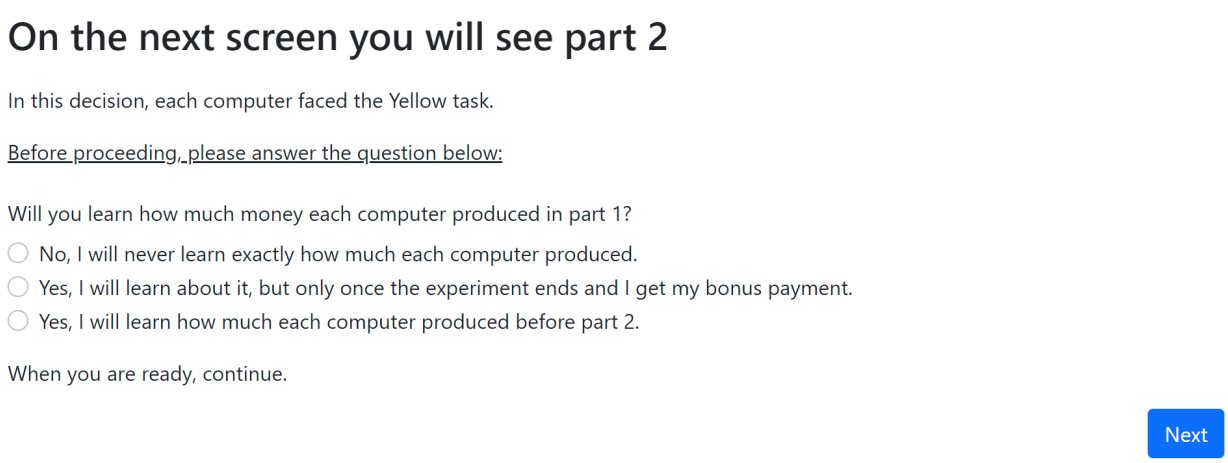


Figure A.19
Note: *Baseline* treatment.

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.

Next

Figure A.20
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.

Review Instructions

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.

Submit

Figure A.21
Note: *Baseline* treatment.

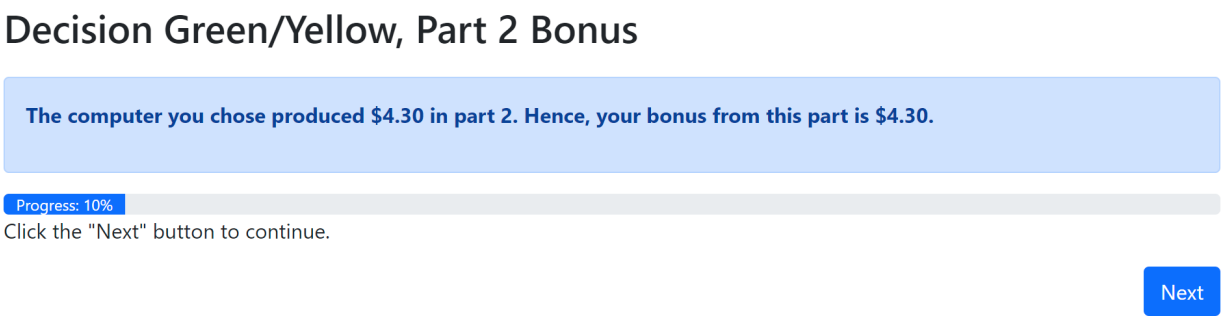


Figure A.22
Note: *Baseline* treatment.

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.

Review Instructions

- **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.

Reminder

- **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.

Submit

Figure A.23
Note: *Strategy Method* treatment.

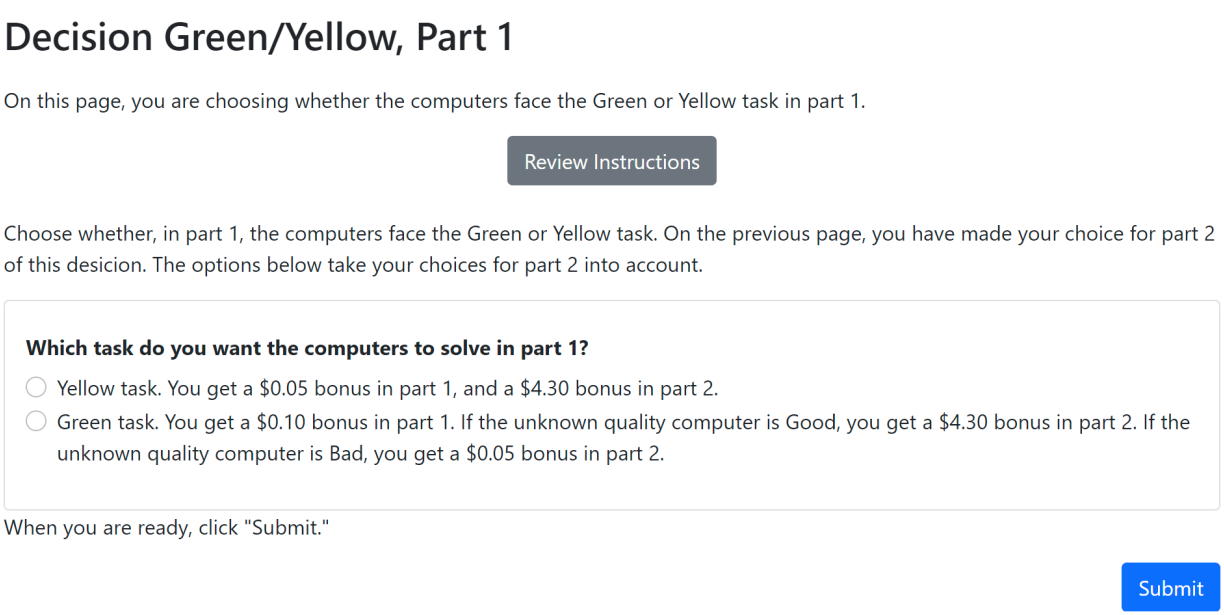


Figure A.24
Note: *Strategy Method* treatment.

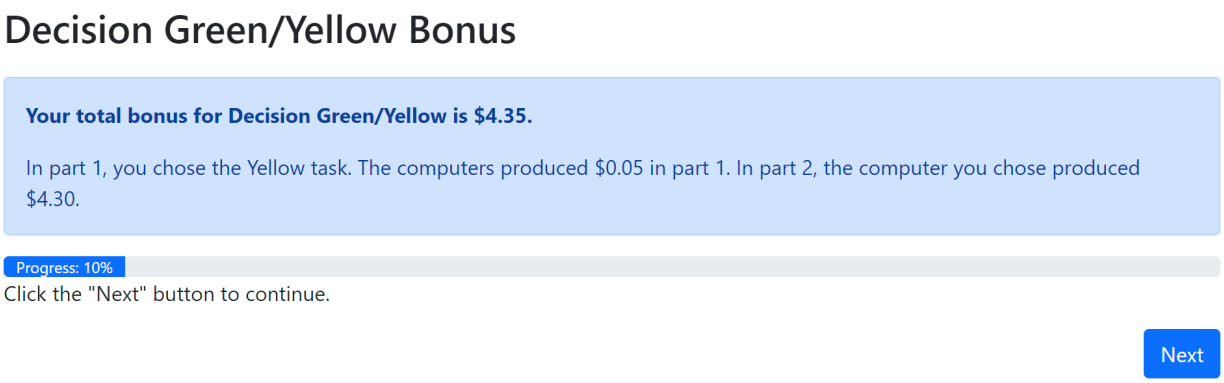


Figure A.25
Note: *Strategy Method* treatment.

Main decision: Automatic Inference treatment

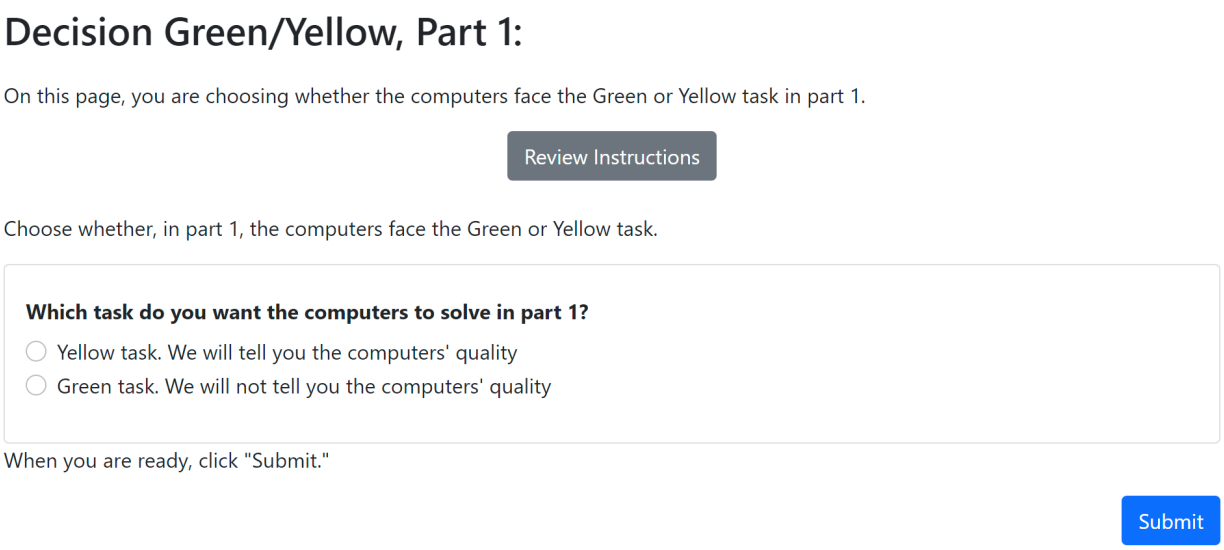


Figure A.26
Note: *Automatic Inference* treatment.

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.

Next

Figure A.27
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.

Review Instructions

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.

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.

Submit

Figure A.28
Note: *Automatic Inference* treatment.

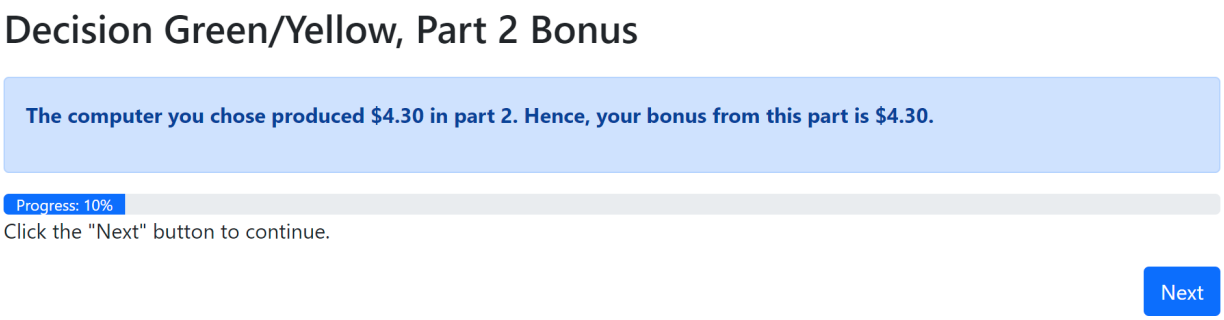


Figure A.29
Note: *Automatic Inference* treatment.

Main decision: Plan treatment

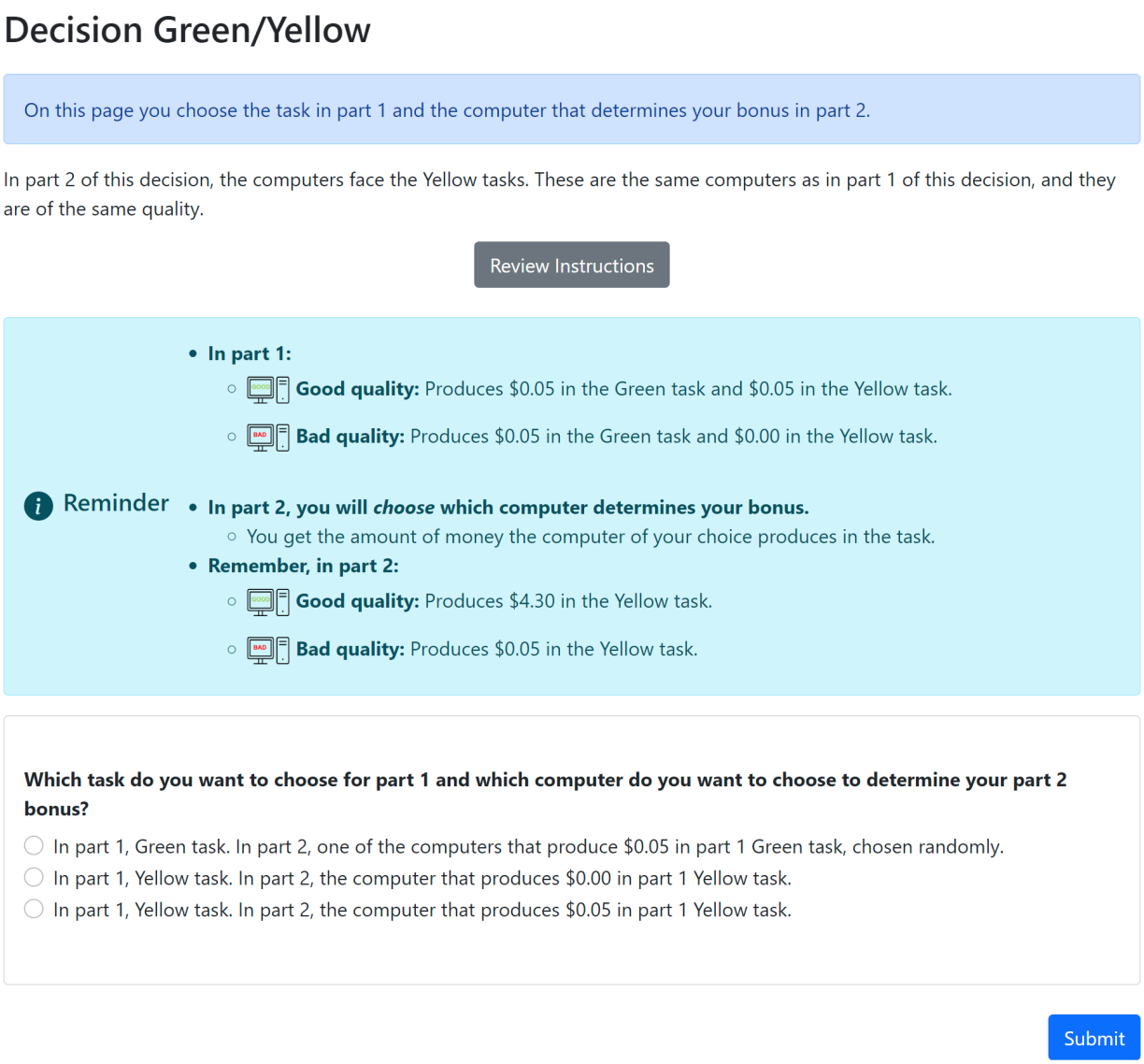


Figure A.30
Note: *Plan* treatment.

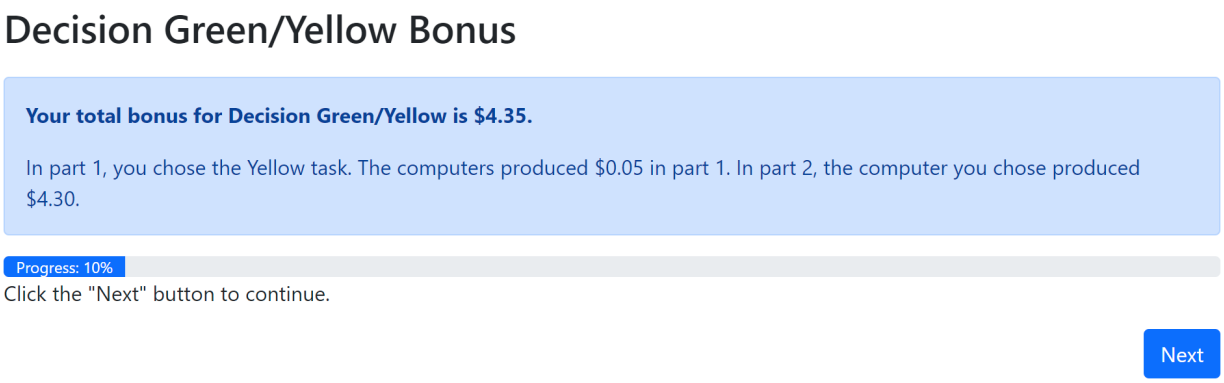
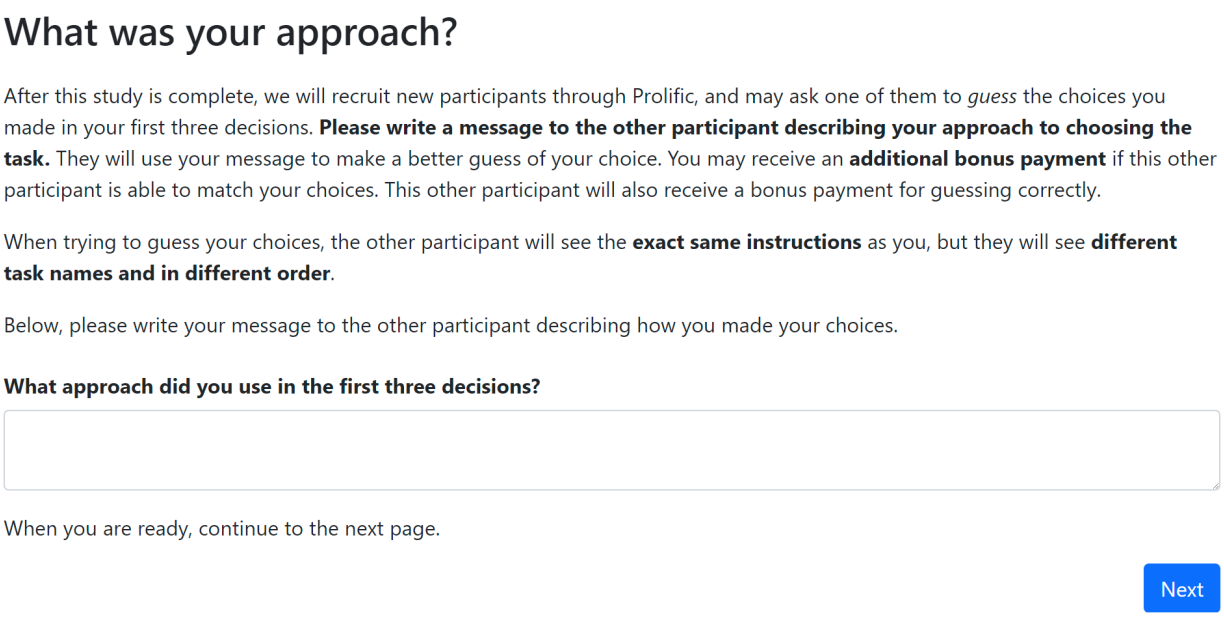


Figure A.31
Note: *Plan* treatment.



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.

Next

Figure A.32
Note: 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.

Next

Figure A.33
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.

Next

Figure A.34
Note: *Plan* treatment.

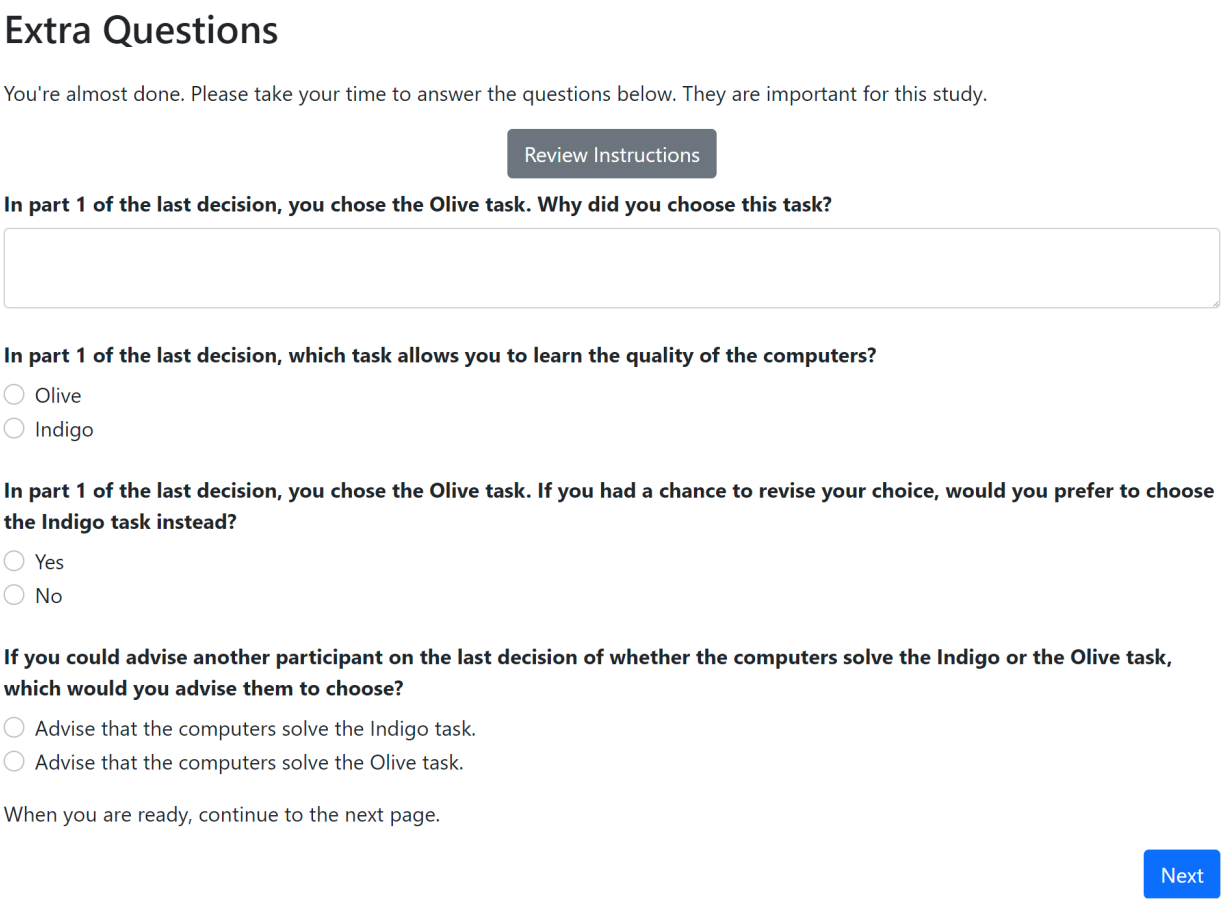


Figure A.35
Note: All treatments.

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.

Next

Figure A.36
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.

Next

Figure A.37
Note: All treatments.

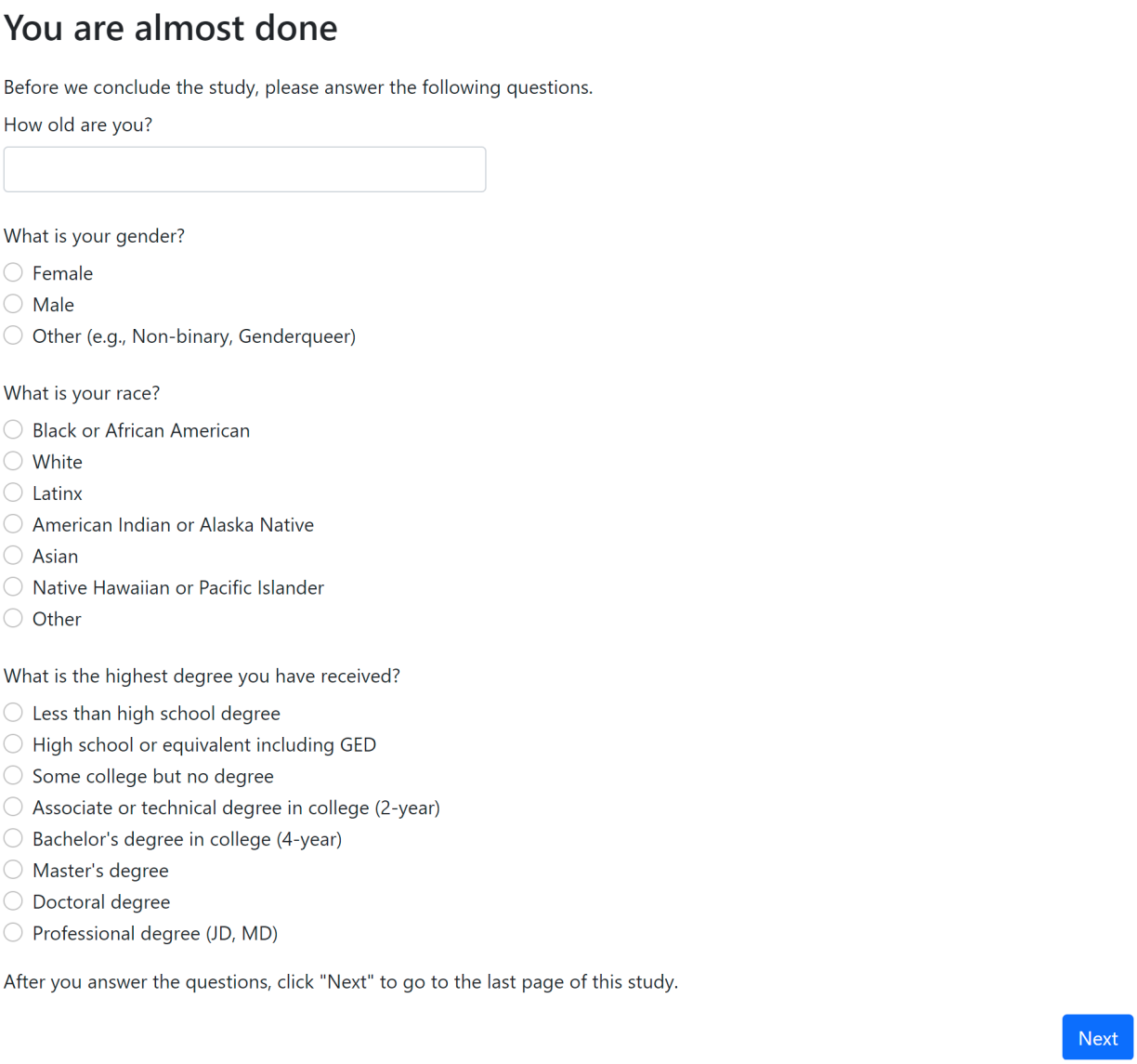


Figure A.38
Note: All treatments.

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.

Submit

Figure A.39
Note: All treatments.